Imperial College London

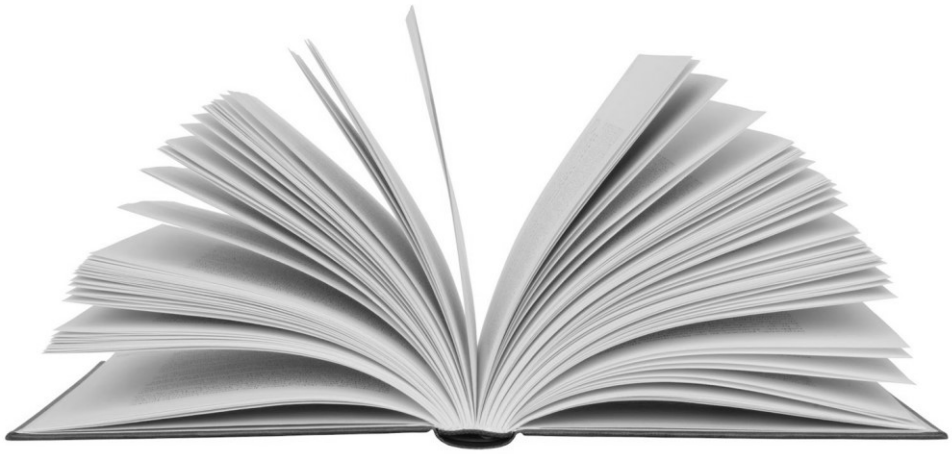
A virtual solution to disparate clinical information systems and consequent data silos

Farzan A. Ramzan

Submitted for the Degree of Doctor of Philosophy

Department of Primary Care & Public Health
School of Public Health
Faculty of Medicine
Imperial College London

[May 2020]



"Each person in the world creates a Book of life.
This book starts with birth and ends with death.
Its pages are made up of the records of the principal events in life.
Record linkage is the name given to the process of
assembling the pages of this Book into a volume".

Dr Halbert Dunn [1]

Abstract

Background: NHS healthcare services in England are fragmented in how they deliver care, and boundaries between primary and secondary care result in the fragmentation of health information. Thus, most healthcare data become trapped in silos within multiple disparate clinical information systems that lack interoperability. Viewing the issue of information fragmentation as an unintended consequence of interactions between health information technologies, this research adopts a sociotechnical stance and identifies informatics techniques that enable the linkage and integration of heterogeneous data silos within an emergency care setting.

Aim: The aim of this research is to investigate and adopt informatics techniques that address health information fragmentation and data silos within an NHS Hospital Emergency Care setting.

Method: This research employed a descriptive observational quantitative single case study design using routinely collected administrative data. The literature identified informatics techniques, incorporating ontologies, clinical data warehouse architectures, data quality indicators and record linkage strategies. A conceptual framework was developed to facilitate the sequential application of these techniques and enable the linkage and integration of multiple heterogeneous datasets to create a linked longitudinal dataset.

Results: Domain understanding via process modelling permitted identification of the extent of fragmentation. Patient ontologies, standardisation approaches and clinical data warehousing enabled interoperability. Probabilistic linkage strategies outperformed deterministic strategies. The creation of a linked dataset facilitated the analysis of departmental transfer times, four-hour standard and interrupted time series analysis comparing emergency admission rates pre/post intervention.

Conclusion: Each integration scenario presents a specific challenge dependent upon the environment in which information fragmentation has occurred. Data integration is not just one problem, and the most important challenge lies in the correct identification and combination of informatics tools, technologies and processes to perform integration. Thus, disparate data can be linked to answer research questions and create new knowledge not previously possible using unlinked data.

Declaration of originality

I declare that the work presented in this thesis is my own. Where information has been derived from other sources, this has been indicated and appropriately referenced.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Funding

I am eternally grateful to the Department of Primary Care & Public Health for providing the opportunity to work the Health Services Research Unit and funding this research.

Acknowledgements

First, I am grateful to God for granting me the strength and perseverance to undertake and complete this research. I would like to express my sincerest gratitude to my mentor and supervisor, Dr Shamini Gnani, for her support and guidance throughout the process, and my co-supervisor Dr Alex Bottle, for his constant advice.

Second, I would also like to extend my thanks to esteemed colleagues Professor Azeem Majeed, Dr Hugh Millington, Dr Medhavi Bucktowonsing, Ms Janet Samuel, Professor Paul Aylin, Professor Alison Holmes, Professor Salman Rawaf, and colleagues within Imperial College Healthcare NHS Trust and the Department of Primary Care & Public Health. In particular, I would like to thank my examiners Professor Ricky Banarsee and Professor Simon de Lusignan, for their valuable feedback and advice.

I have had the privilege of making some brilliant friends and would like to express my gratitude to Dr Eugenio Macchiarulo, Dr Bushra Siddiqui, Dr Amanda Mathew, Dr John Tayu Lee, Dr Michele Davison, Dr Elizabeth Dubois, Dr Richard Patterson, Dr Philip Anyanwu, Dr Mohammed Alzamil, Dr Mohammed Aljuaid, Dr Elizabeth Koshy, Dr Anthony Thomas, and many others not mentioned.

My acknowledgements would not be complete without thanking Professor Natalie Shlomo and Dr Katie Harron for their insightful record linkage courses and practical advice at the early stages of this research.

Most of all, a special thank you to my wonderful family, to whom I owe a debt of gratitude for their unconditional support, encouragement, and patience throughout my studies, without which this work would not have been possible.

Finally, I would like to dedicate this work to the memory of those that were present at the start of this journey, but sadly departed this world before its completion, may God rest their souls.

Table of Contents

Abstract 3
Table of Contents 6
List of tables 11
List of figures 13
Abbreviations 15
Chapter 1 Introduction 16
1.1 Overview 17
1.2 Problem statement 17
1.3 Rationale for research 19
1.4 Research Questions 21
1.5 Research Aim and Objectives 21
1.6 Research Hypothesis 21
1.7 Thesis outline 22
1.8 Research outputs from this PhD 23
Chapter 2 Literature Review 24
2.1 Introduction 25
2.2 Co-location of Primary Care with ED 25
2.2.1 The National Health Service 25
2.2.2 Challenges facing the NHS 28
2.2.3 Emergency Medicine 29
2.2.4 Accident & Emergency 29
2.2.5 A&E Initial Assessment Triage Category 31
2.2.6 A&E Stream 33
2.2.7 Clinical Quality Indicators 34
2.2.8 The four-hour standard 34
2.2.9 National statistics 37
2.2.10 Overcrowding within EDs 39
2.2.11 Emergency admissions 43
2.2.12 Patient preference for ED 45
2.2.13 Government urgent care service initiatives 48
2.2.14 Primary Care in ED 50
2.2.15 Section summary 56
2.3 Data fragmentation 56
2.3.1 Health information fragmentation 56

2.3.2 Information fragmentation within emergency medicine 60
2.3.3 Medical records 62
2.3.4 Clinical information systems 66
2.3.5 Disparate clinical information systems 71
2.3.6 Interoperability 73
2.3.7 Medical informatics 76
2.3.8 Section summary 78
2.4 Data integration 79
2.4.1 Ontologies 82
2.4.2 Health Level 7 91
2.4.3 Knowledge discovery in databases 94
2.4.4 Clinical data warehouse 96
2.4.5 Section summary 101
2.5 Data quality 102
2.5.1 Data quality as 'fit for purpose' 104
2.5.2 NHS number validation 108
2.5.3 Data quality management 109
2.5.4 Section summary 111
2.6 Data linkage 112
2.6.1 Deterministic Record Linkage 113
2.6.2 DRL strategy: Hospital Episode Statistics 115
2.6.3 Probabilistic Record Linkage 116
2.6.4 Complementary techniques to enhance RL 121
2.6.5 Linkage strategy 125
2.6.6 Blocking strategy 125
2.6.7 Soundex algorithm 126
2.6.8 Combination ID 127
2.6.9 Split Date of birth 128
2.6.10 Relative Frequency 128
2.6.11 String comparator 129
2.6.12 Expectation Maximisation algorithm 130
2.6.13 RL Software selection 131
2.6.14 Application of RL strategies 131
2.6.15 Section summary 132
2.7 Chapter summary 133
Chapter 3 Research Framework & Methodology 134

3.1 Introduction 135
3.2 Theoretical framework 135
3.2.1 Systems Theory 135
3.2.2 GST & Information system performance 137
3.2.3 Sociotechnical Systems Theory 139
3.2.4 Interactive Sociotechnical Analysis framework 141
3.2.5 The relevance of STS to this research 144
3.3 Conceptual framework 146
3.3.1 The conceptual framework for this research 146
3.4 Research methodology 149
3.4.1 Epistemology 149
3.4.2 Philosophy 150
3.4.3 Approach to theory development 151
3.4.4 Research design 153
3.4.5 Data collection 154
3.4.6 Ethics 156
3.4.7 Research strategy 157
3.4.8 Importance of undertaking this research 159
3.5 Research setting 161
3.5.1 The London Borough of Hammersmith & Fulham 161
3.5.2 Healthcare services within H&F 162
3.5.3 Urgent and Emergency Care 162
3.5.4 Clinical Information Systems 172
3.6 Chapter summary 174
Chapter 4 Methods 175
4.1 Introduction 176
4.2 Domain understanding 176
4.2.1 Process modelling 176
4.2.2 Training 177
4.3 Semantic interoperability 178
4.3.1 Ontologies 178
4.3.2 Standardisation of data elements 178
4.4 Syntactic interoperability 179
4.4.1 Clinical Data Warehouse 179
4.4.2 EMR record structure 180
4.5 Data quality 181

4.5.1 Data quality assessment 181
4.6 Data linkage 186
4.6.1 Deterministic linkage strategy 187
4.6.2 Hybrid (DRL & PRL) linkage strategy 187
4.7 Linked dataset analysis 196
4.7.1 Transition times 196
4.7.2 ED four-hour standard 197
4.7.3 Comparison of admission rate pre/post UCC 198
Chapter 5 Results 201
5.1 Introduction 202
5.2 Domain understanding 202
5.2.1 Activity diagram 203
5.2.2 Use case diagram 207
5.2.3 Heterogeneity amongst CIS 209
5.3 Semantic interoperability 211
5.3.1 Ontologies 211
5.3.2 Standardisation of data elements 212
5.4 Syntactic interoperability 214
5.5 Data quality 216
5.5.1 Data quality assessment 216
5.6 Data linkage 226
5.6.1 Hybrid (DRL & PRL) linkage 226
5.6.2 Comparison of linkage strategies 235
5.6.3 Linked dataset 237
5.7 Linked dataset analysis 239
5.7.1 Transition times 239
5.7.2 ED four-hour standard 243
5.7.3 Comparison of admission rate pre/post UCC 248
Chapter 6 Discussion 254
6.1 Introduction 255
6.2 Summary of the study 255
6.2.1 Problem 255
6.2.2 Research approach 255
6.3 Discussion of the results 257
6.3.1 Selection of informatics techniques 257
6.3.2 Domain understanding 259

6.3.3 Semantic interoperability 261
6.3.4 Syntactic interoperability 263
6.3.5 Data quality 264
6.3.6 Data linkage 266
6.3.7 Linked dataset analysis 270
Chapter 7 Conclusion 273
7.1 Conclusion 274
7.1.1 Problem classification (Figure 7.1) 274
7.1.2 Sequential process (Figure 7.1) 276
7.1.3 Outcome (Figure 7.1) 277
7.2 Concluding comments 277
7.3 Implications 279
7.4 Contribution to knowledge 280
7.4.1 Informatics techniques 280
7.4.2 Record linkage strategy 280
7.4.3 Linked dataset analysis 281
7.5 Limitations 282
7.5.1 Level of interoperability 282
7.5.2 Level of heterogeneity 282
7.5.3 NHS number data quality 283
7.5.4 Study setting 284
7.6 Recommendations for future work 284
7.6.1 CDW linked dataset development 284
7.6.2 Data linkage 285
7.6.3 Primary care collaboration 285
References 287

List of tables

Table 2.1 NHS England new care models 27
Table 2.2 Types of Accident and Emergency departments 30
Table 2.3 A&E initial assessment triage category 32
Table 2.4 Overview of A&E clinical quality indicators 36
Table 2.5 Type 1 ED attendances, admissions and four-hour target in England 37
Table 2.6 Examples of primary care involvement in ED 52
Table 2.7 Layers of interoperability (adapted from Benson et al., 2016) 74
Table 2.8 Common difficulties relating to interoperability (adapted from Benson et al., 2016) 75
Table 2.9 Different types of system interrelationships, adapted from Das, 1992 81
Table 2.10 Types of ontologies (adapted from Guarino 1998) 86
Table 2.11 Aspects of terminologies similar to languages (adapted from Rector et al. 2014) 86
Table 2.12 Types of integration facilitated by ontologies (adapted from Liaw et al., 2014) 88
Table 2.13 Comparison of semantic interoperability and ontological approach (adapted from Liyanage et al. 2015) 88
Table 2.14 Semantic conflicts (adapted from Gagnon, 2007). 89
Table 2.15 Ontology-based integration approach (adapted from Wache et al, 2001). 90
Table 2.16 HL7 layers (adapted from Benson et al., 2016) 91
Table 2.17 HL7 FHIR Detailed description of Resource: Patient, Element: name 93
Table 2.18 Comparison of online transaction processing and data warehouse systems, adapted from Reeves, 2013. 98
Table 2.19 Data quality dimensions and associated calculations 106
Table 2.20 Example of deterministic record linkage 114
Table 2.21 Step based approach for HES matching criteria 116
Table 2.22 Simplified overview of Fellegi & Sunter PRL model. 117
Table 2.23 Comparison of classification pairs, adapted from Sayers (2015) and Blakely (2002). 119
Table 2.24 Research articles applying DRL and PRL techniques to demographic attributes. 122
Table 2.25 An overview of common linkage strategies and techniques per author. 124
Table 3.1 Types of interaction within the ISTA framework (adapted from Harrison (2007) and Booth (2015)) 143
Table 3.2 Overview of the methodological approach for this research. 149
Table 3.3 Overview of streaming designations (adapted from Gnani et al., 2013) 166
Table 3.4 Staffing per site pre/post UCC 169
Table 3.5 Overview of departmental Clinical Information Systems 173
Table 4.1 M and U probabilities used within the PRL approach (Step 19) 193
Table 5.1 Summary of information related factors identified along the urgent & emergency care pathway (Figure 5.1) 204
Table 5.2 Interactions between actors, based on Use case diagram (Figure 5.2) 208
Table 5.3 Heterogeneity amongst urgent & emergency care CIS 209
Table 5.4 Person attributes, adapted from NHS Digital (2019). 211
Table 5.5 Standardisation of naming conventions 213
Table 5.6 Episode ID data quality summary 216
Table 5.7 Patient ID data quality summary 217
Table 5.8 NHS number data quality summary 218
Table 5.9 Patient surname data quality summary 219
Table 5.10 Patient forename data quality summary 219

Table 5.11 Patient sex data quality summary 221
Table 5.12 Patient date of birth data quality summary 221
Table 5.13 Patient postcode data quality summary 222
Table 5.14 Episode arrival date/time data quality summary 223
Table 5.15 Episode depart date/time data quality summary 224
Table 5.16 Overview of unique record pre/post DRL deduplication. 230
Table 5.17 Comparison of linkage approaches 236
Table 5.18 Example of Transition time calculation 239
Table 5.19 Chronological discrepancies between services. 243
Table 5.20 Comparison of results from Linked and Unlinked datasets 246
Table 5.21 Baseline socio-demographic characteristics of ED attendances and hospital admissions 252
Table 5.22 Effect of UCCs on total attendances by hospital site 253
Table 5.23 Adjusted multivariate model of effect size of UCCs on emergency hospital admission by hospital site 253

List of figures

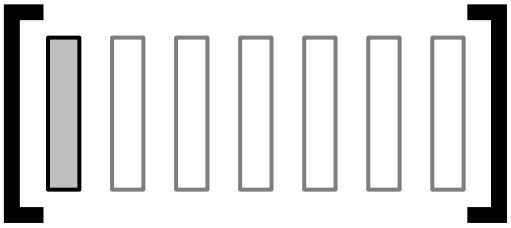
Figure 2.1 NHS funding flow (adapted from an infographic by Will Stahl-Timmins http://www.bmj.com/content/356/bmj.j41) 26
Figure 2.2 Flow of patients in an emergency department (adapted from Fatovich et al. 2002) 31
Figure 2.3 Monthly attendance at Type 1 Major A&E departments in England. 38
Figure 2.4 Percentage of attendances with a total time in department of greater than four hours, for Type 1 Major A&E departments in England. 38
Figure 2.5 Monthly emergency admissions via Type 1 Major A&E departments in England. 38
Figure 2.6 Simplified example of the relational model 68
Figure 2.7 Data entry and storage within clinical information systems 69
Figure 2.8 Hierarchical MeSH tree structure for term Alzheimer Disease 83
Figure 2.9 Ontologies populated by concepts (adapted from Jakus et al., 2013) 84
Figure 2.10 Types of ontologies (adapted from Guarino 1998) 85
Figure 2.11 Organisation of concepts, attributes and relations (adapted from Jakus et al., 2013) 87
Figure 2.12 Example of XML with patient details 92
Figure 2.13 The data pipeline, adapted from Burton et al. (2015). 95
Figure 2.14 Basic data warehouse environment (adapted from Reeves, 2013). 98
Figure 2.15 Comparison of Inmon and Kimball DW architectures (adapted from Reeves, 2013). 99
Figure 2.16 Challenges for healthcare data (adapted from Iyamu et al. (2018) 103
Figure 2.17 Conceptual framework of data quality (adapted from Wang et al., 1996) 105
Figure 2.18 Validation process for the NHS number 108
Figure 2.19 Data quality functions (adapted from Davoudi et al., 2015) 109
Figure 2.20 Path of hospital data extracts into the HES database 115
Figure 2.21 Thresholds used in probabilistic record linkage. 121
Figure 2.22 overview of the Jaro algorithm. 129
Figure 2.23 Interconnection of chapter sub-sections 133
Figure 3.1 System interconnection and interdependence (adapted from Teater (2014)). 136
Figure 3.2 The Web of System Performance (adapted from Whitworth et. at (2008) 138
Figure 3.3 Interactive sociotechnical systems, adapted from Harrison et al. (2007). 142
Figure 3.4 Conceptual framework for this thesis 147
Figure 3.5 Population pyramid of England and Hammersmith & Fulham, 2015 (source Office of National Statistics). 161
Figure 3.6 Emergency Department pathway pre-UCC 164
Figure 3.7 Patient pathway across the integrated GP-led model of care (Gnani et al., 2013) 165
Figure 3.8 Episodes of attendance and admissions throughout the study period at HH site 167
Figure 3.9 Episodes of attendance and admissions throughout the study period at CXH site 168
Figure 3.10 HH site patient population pyramid pre/post UCC 170
Figure 3.11 CXH site patient population pyramid pre/post UCC 171
Figure 3.12 Timeline of departmental clinical information system implementation. 173
Figure 4.1 Data element cleaning via the use of reference tables 182
Figure 4.2 Incorporation of Quasi-identifier. 188
Figure 4.3 Reference tables for each ID type data element. 189
Figure 4.4 Example of identifying the number of Surname variants per unique ID 190
Figure 4.5 Triangulation of a patient within distinct IDs. 191
Figure 4.6 Application of probabilistic record linkage 194

Figure 4.7 Creation of Linked dataset via linkage of UCC, ED and IPC datasets with Master Patient Table.....196
Figure 4.8 Interrupted time series design and regression analysis (Gnani et al., 2013).....199
Figure 5.1 UML Activity diagram of information capture via urgent & emergency care patient pathway. 203
Figure 5.2 Use case diagram highlighting databases encountered during patient episode.....207
Figure 5.3 Patient ontology model 212
Figure 5.4 Data extraction and import into the clinical data warehouse 214
Figure 5.5 Data quality dimension: Completeness for selected data elements 225
Figure 5.6 Hidden complex interrelationship between relative CIS IDs for a unique patient within each dataset.....226
Figure 5.7 Variants of data elements per Pseudo ID, NHS number and Quasi-ID 228
Figure 5.8 Deduplication of patient identification numbers. 229
Figure 5.9 Distribution of match weights based on Soundex Surname.....231
Figure 5.10 Distribution of match weights based on Soundex Forename.....232
Figure 5.11 Soundex Surname & Soundex Forename matching data elements 234
Figure 5.12 Comparison of HES algorithm with the Hybrid algorithm 235
Figure 5.13 Structure of patient records and associated episodes records within the Linked dataset. 238
Figure 5.14 Boxplot of transition times from UCC referral to ED across CXH and HH sites. 240
Figure 5.15 Boxplot of referral times from ED to IPC across CXH and HH sites.....241
Figure 5.16 Boxplot of referral times from ED to UCC, across CXH and HH sites. 242
Figure 5.17 HH ED comparison of time in ED within Linked and Unlinked datasets 244
Figure 5.18 CXH ED comparison of time in ED within Linked and Unlinked datasets 245
Figure 5.19 Episodes of attendance and admissions pre/post UCC launch at HH site.....248
Figure 5.20 Episodes of attendance and admissions pre/post UCC launch at CXH site 250
Figure 7.1 Conceptual conclusion based on research findings.....275

Abbreviations

Abbreviation	Term
A&E	Accident & Emergency
API	Application Program Interfaces
BPM	Business Process Modelling
CCGs	Clinical commissioning groups
CDS	Commissioning Data Set
CDW	Clinical Data Warehouse
CIS	Clinical Information System
CQI	Clinical Quality Indicators
CXH	Charing Cross Hospital
DM	Data Mining
DQ	Data Quality
DRL	Deterministic Record Linkage
ED	Emergency Department
EHR	Electronic Health Record
EM	Emergency Medicine
EMR	Electronic Medical Record
EPR	Electronic Patient Record
GP	General Practitioner
GST	General Systems Theory
H&F	Hammersmith & Fulham
HES	Hospital Episode Statistics
HH	Hammersmith Hospital
HIE	Health Information Exchange
HIT	Healthcare Information Technology
HL7	Health Level 7
ICHNT	Imperial College Healthcare NHS Trust
ICU	Intensive Care Unit
IPC	Inpatient Care
IS	Information Systems
ISTA	Interactive Sociotechnical Analysis
ITS	Interrupted Times Series
KDD	Knowledge Discovery in Databases
MPI	Master Patient Index
NHS	National Health Service
NHS-PDS	NHS personal demographics service
PAS	Patient Administration System
PDS	Personal Demographics Service
PID	Patient Identification number
PRL	Probabilistic record linkage
QID	Quasi-Identifier
RDBMS	Relational Database Management System
RL	Record linkage
SQL	Structured Query Language
STS	Sociotechnical Systems
UCC	GP-led Urgent Care Centre
UK	United Kingdom of Great Britain & Northern Ireland
UML	Unified modelling language
XML	Extensible Mark-up Language

Chapter 1 Introduction



1.1 Overview

This chapter introduces the research undertaken within this thesis. It begins with a description of the research problem and specifies the question to be addressed and motivation for conducting this research. It then provides the aims, objectives and hypothesis to be tested, finally providing a brief overview of remaining chapters. Moreover, following a review of the literature, the problem statement, aims and objectives were revised to reflect this research more accurately.

1.2 Problem statement

The fragmentation of health information is prevalent throughout NHS healthcare services in England. Although the NHS is the largest semi-integrated health system in the world, it remains fragmented in how it organises the delivery of care, hence, boundaries between the provision of primary and secondary care will be inevitable. [2]

A 'silo' is a system, process or department that operates in isolation from others. [3] Organisational silos inadvertently develop amongst services via acquired specialty expertise or structural configuration. [4] Consequently, most healthcare data are trapped in silos across health systems, e.g. between primary and secondary care, or even within the same hospital setting, between different departments. [5]

Healthcare organisations are information-intensive enterprises, and clinical personnel require data and information to make appropriate decisions to provide patient care. [6] Within hospital settings, there are many departments and clinical specialties that need to exchange information. [7]

However, there is no 'one' overall Clinical Information System (CIS) designed to offer access to all the available information within a hospital due to the complexity of patient information. Hence, an individual department will typically possess bespoke CISs that record information specific to the functionality of the department. [8]

Moreover, CISs are designed to store, process and exchange medical information [9] to facilitate patient care [10] by housing Electronic Medical Records (EMR), i.e. all data and documents generated or received during the care of a patient at a healthcare institution. [9]

Thus, within one hospital setting, patient information may be difficult to access [11] as medical data in various formats is dispersed amongst multiple proprietary disparate departmental CIS [12] that comprise heterogeneous hardware platforms, operating systems, network protocols, [13] and relational data models, each functioning as an independent unit. [14] For example, the CISs within ambulatory, inpatient, and emergency settings. [11]

Consequently, EMRs concerning the same patient but transcribed within different CIS and allocated unrelated identifiers will inevitably result in fragmentation of the patient record, and lead to future complications when trying to re-identify the same patient. Inevitably, resulting in the formation of information 'islands' or data silos, that exist/remain in isolation from other departments. [5]

Data silos arise due to individual departments within an organisation using more than one CIS designed by different vendors, to manage patients. However, these heterogeneous CIS do not 'talk' to one another and become isolated [15] due to a lack of interoperability, which is defined as the ability of two or more systems or components to exchange information and to use the information that has been exchanged. [16]

Nonetheless, the fragmentation of patient information risks patient safety; hinders data aggregation for clinical decision support; prevents physicians from having comprehensive medical information; deters effective population health approaches; creates inefficiencies by delaying care; limits public health reporting; and reduces the utility of electronic data for clinical research. [17]

Clinical decision-making relies upon the availability of information to inform communication, [18] failure to communicate effectively may lead to errors or result in adverse clinical events or outcomes, e.g. clinical morbidity or mortality. [19, 20] Therefore, as every clinical decision is predicated on information about the patient and treatment options available, the quality of information may inadvertently affect the level of care a patient receives. [21, 22]

As the amount of heterogeneous medical data stored within different subsystems increases, [15] so too does the potential for errors, inconsistencies and duplication of this data. Therefore, to support work within hospitals, it is necessary to integrate these subsystems. However, integrated systems that satisfy all requirements on information processing within hospitals are not easily available. [8] Thus, healthcare organisations and hospitals face many challenges such as technical, security and human interaction in freeing data from silos. [15]

Additional factors contributing to disparities include; first, the heterogeneity of CIS or platforms hosting data [23], e.g. multi-vendor modern/legacy systems [12] as these determine the 'knowledge' possessed by each CIS and the extent of common ground between these systems. [24] Second, the heterogeneity of data at the semantic (interpreting meaning), syntactic (formatting), and structural (storing) levels. [23] Third, data complexity, noise and incompleteness, [15, 23] all of which act as barriers to interoperability and leave medical information locked in silos within CISs. [12]

Moreover, NHS England describes information as its lifeblood, a vital resource for the clinical management of individual patients and the efficient management of services. [25] Healthcare is an information-based science, and much of clinical practice involves the gathering, synthesising and acting upon information. [26] Thus, facilitating data sharing by addressing interoperability is the only sustainable way for enabling CIS to talk to one another in order to obtain the complete 'image' of a patient, to support clinical patient care and medical research. [15]

However, from a theoretical perspective, there is a lack of understanding in terms of what fundamental processes are capable of integrating information, and whether or not there is a unifying theory of integration that incorporates the separate challenges of data integration and information exchange. [27]

Consequently, the data silo problem remains salient and unsolved within healthcare environments, [28] and interoperability remains a key issue and a long-term domain of research for healthcare information systems. [29] The fragmentation of information underlies the healthcare crisis, [30] and investigation into addressing the problem of information fragmentation is of crucial importance.

1.3 Rationale for research

This thesis presents an original perspective on the phenomenon of Health Information Fragmentation (HIF) and argues that although a unifying theory of data integration/exchange [27] has yet to be established, the issue of HIF can, nevertheless, be addressed by using selective informatics techniques identified within the research literature.

Since the data silo problem is sociotechnical [28], this research, underpinned by Sociotechnical Systems Theory, investigates how the implementation of new Health Information Technology results in the unintended consequence and emergence of the invisible problem of HIF within an organisation. [30-32]

Therefore, to permit an in-depth study of this phenomenon, an Observational Quantitative Single Case Study Research Design was employed using historical routinely collected administrative data, spanning several years. This study provides a real-world example of HIF caused by data held in silos within multiple disparate CISs that lack any common identifiers, housed within an Urgent Care Centre, Emergency Department and Inpatient Care Service of a large metropolitan London hospital.

Furthermore, the creation of a new dataset does not necessarily constitute new knowledge. Data must first be analysed and interpreted before it can be said to have yielded fresh insights and understanding. [33] Therefore, this thesis presents three examples of analyses that utilise the linked dataset to provide proof of concept and demonstrate its practical applicability.

In summary, this research:

- Presents appropriate background to understanding the factors and causes leading to the formation of data silos within a hospital setting.
- Identifies fundamental processes, theories and methodologies from within the research literature that facilitate the standardisation, integration and linkage of disparate data.
- Develops an appropriate conceptual framework, end-to-end, from initial data silo identification to the creation of a linked dataset and evidences this via practical application.
- Demonstrates the benefits and practical applicability of the linked longitudinal dataset by presenting analyses that would not have been possible using unlinked datasets.

Thus, this research takes the reader from the beginning stage of investigating data silo formation, to the end-stage that results in the creation of a linked longitudinal dataset, used thereafter as the source for specific analyses to demonstrate practical application.

1.4 Research Questions

1. What informatics techniques address the issue of health information fragmentation?
2. How can identified informatics techniques be adopted to support the linkage and integration of disparate data silos within a hospital setting?

1.5 Research Aim and Objectives

The overall aim of this research is to adopt informatics techniques that address data silos within an NHS Hospital Emergency Care setting.

Research objectives

1. To explore the formation of data silos within healthcare settings.
2. To identify informatics techniques that enable the linkage and integration of heterogeneous data sources.
3. To apply techniques that enable interoperability between the clinical information systems in the emergency care setting.
4. To select data warehouse architecture appropriate for housing healthcare data extracts.
5. To determine if data are fit for linkage purposes by assessing the quality of data.
6. To differentiate the effect of deterministic and non-deterministic data linkage strategies with respect to the numbers of correctly matched records.
7. To calculate the transition time for patient transfer between urgent and emergency care services.
8. To distinguish the effect of utilising Linked and Unlinked datasets upon the ED 4 hour standard.
9. To compare the rate of emergency patient admissions pre/post UCC implementation.

1.6 Research Hypothesis

A review of the literature focusing on record linkage theory (page 128) led this study to speculate that employing either deterministic or probabilistic linkage strategies would yield different results. Objective six sought to differentiate this assumption empirically, i.e. whether the selection and application of a particular linkage strategy influenced the number of correctly identified records belonging to the same individual across multiple datasets, stating the hypothesis as:

H_0: *There is no difference between using deterministic or probabilistic record linkage techniques.*
H_1: *There is a difference between using deterministic or probabilistic record linkage techniques.*

1.7 Thesis outline

- **Chapter 1 Introduction:** provides a general overview of this research, providing the rationale, highlighting the problem and specifying research questions, aims, objectives and hypothesis that guide the study.

- **Chapter 2 Literature Review:** sets the background for the study setting in which information fragmentation has occurred, and contextualises the research problems, detailing informatics approaches that facilitate data linkage and integration.

- **Chapter 3 Research Framework & Methodology:** presents the theoretical framework adopted in this study, and the conceptual framework developed to facilitate linkage and integration. In addition, the research methodology and philosophical underpinning are presented.

- **Chapter 4 Methods, Chapter 5 Results and Chapter 6 Discussion:** To address the research questions and objectives of this study, each chapter address the following sections sequentially: Domain understanding, Semantic interoperability, Syntactic interoperability, Data quality, Data linkage and Linked data analysis.

- **Chapter 7 Conclusion:** draws together the findings from this research, provides a conclusion for the entire study, discusses the limitation and presents recommendations.

1.8 Research outputs from this PhD

Original peer-reviewed publications directly related to this work

1. Gnani S, **Ramzan F**, Ladbrooke T, Millington H, Islam S, Car J, et al.
   Evaluation of a general practitioner-led urgent care centre in an urban setting: description of service model and plan of analysis. JRSM short reports. 2013. [34]

2. Cowling TE, **Ramzan F**, Majeed A, Gnani S. [Rapid response]
   Attendances at Charing Cross and Hammersmith Hospitals' urgent care centres, 2009-12. BMJ (Clinical research ed). 2013. [35]

3. Cowling TE, **Ramzan F**, Majeed A, Gnani S.
   Data from two London hospitals' urgent care centres, 2009-12. BMJ: British Medical Journal, 2013. 347(7935): p. 24-24. [36]

4. Amiel C, Williams B, **Ramzan F**, Islam S, Ladbrooke T, Majeed A, et al.
   Reasons for attending an urban urgent care centre with minor illness: a questionnaire study. Emergency medicine journal: EMJ. 2014. [37]

5. Gnani S, McDonald H, Islam S, **Ramzan F**, Davison M, Ladbrooke T, et al.
   Patterns of healthcare use among adolescents attending an urban general practitioner-led urgent care centre. Emergency medicine journal: EMJ. 2014. [38]

6. Cowling TE, **Ramzan F**, Ladbrooke T, Millington H, Majeed A, Gnani S.
   Referral outcomes of attendances at general practitioner led urgent care centres in London, England: retrospective analysis of hospital administrative data. Emergency medicine journal: EMJ. 2015. [39]

7. Gnani S, Morton S, **Ramzan F**, Davison M, Ladbrooke T, Majeed A, et al.
   Healthcare use among preschool children attending GP-led urgent care centres: a descriptive, observational study. BMJ Open. 2016. [40]

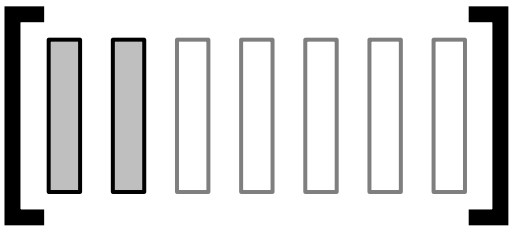
Award for best poster presentation

- Royal College of General Practitioners City Health Conference - London (April 2013)
  Reasons for attending an urban urgent care centre with minor illness: a questionnaire study

Original peer reviewed publication (not directly related to this work)

8. Alshaikh F, **Ramzan F**, Rawaf S, Majeed A.
   Social network sites as a mode to collect health data: a systematic review. Journal of medical Internet research. 2014. [41]

Chapter 2 Literature Review



2.1 Introduction

This chapter contextualises this research by providing a review of the literature, highlighting the main issues, methodologies, theories and developments in the field of health informatics relevant to this research topic as a whole.

Furthermore, the literature review helped to emphasise the importance of this research and provided direction for the research questions and hypothesis, [42] and led to the identification of the research gap. [43] This chapter has been categorised under the following five themes:

1. Co-location of Primary Care with ED
2. Data fragmentation
3. Data integration
4. Data quality
5. Data linkage

2.2 Co-location of Primary Care with ED

2.2.1 The National Health Service

The National Health Service (NHS) was launched in 1948 and was based on the principle that good healthcare should be available to all, regardless of wealth. It embodies three core principles: firstly, that it meets the needs of everyone; secondly, be free at the point of delivery; and lastly, be based on clinical need, not ability to pay; hence, it remains free at the point of use for all United Kingdom of Great Britain & Northern Ireland (UK) residents. [44]

The NHS in England is the largest and most complex health system in the UK. Whilst it is able to provide treatment for the majority of the population, inevitably, it does consume a large proportion of government funding. For example, secondary care services consume £71.9 billion, which is 44 per cent of total government funding. Such services include Urgent & Emergency Care, Mental health, Community health, Planned hospital care, Rehabilitative care, NHS 111 and Learning disability services. [45]

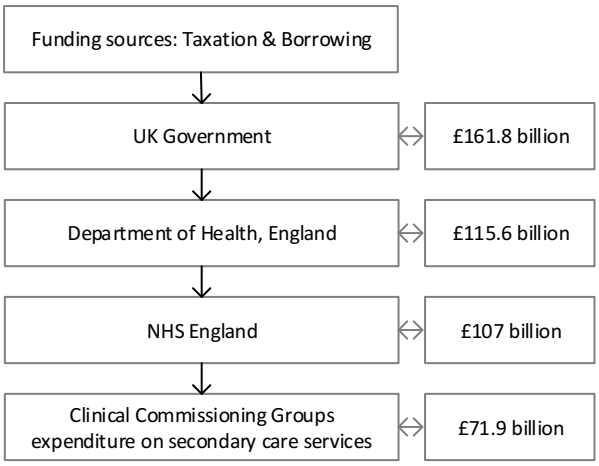


Figure 2.1 NHS funding flow (adapted from an infographic by Will Stahl-Timmins http://www.bmj.com/content/356/bmj.j41)

The NHS is divided into Primary & Community care, Secondary care, and Tertiary care. For people experiencing unfamiliar symptoms, suffering from an illness or injury, the first point of contact is usually Primary care, which is provided by healthcare professionals that include General Practitioners (GPs), dentists, pharmacists and optometrists. Primary care can also be thought of as a 'gateway' to receiving more specialist care and works closely with community and social care.

In some cases, patients who have initially been seen by primary care professionals may be referred to a secondary care professional, or a specialist with expertise on the patient's issue. Secondary care is provided by consultant-led services, e.g. psychology, psychiatry and orthopaedics. As it is delivered within a hospital environment, it is sometimes referred to as 'hospital care' or 'unscheduled care' and can either be planned (elective), e.g. cataract operation, or unplanned (non-elective) urgent and emergency care, e.g. treatment for a fracture.

Patients that are subsequently hospitalised may require highly specialised treatment and care. Tertiary care refers to highly specialised treatment by professionals, e.g. surgeons with specific expertise in a given field, to carry out investigation and treatment for the patient, examples include neurosurgery, cardiac surgery, cancer management, transplants and secure forensic mental health services. [46]

The main role for NHS England is to set the priorities and direction of the NHS and to improve health and care outcomes for people in England, and as such, it is the commissioner for primary care services, military health services and some specialised services. NHS England manages approximately

£107 billion of the overall NHS budget and ensures that organisations are spending allocated funds effectively, with resources being allocated to Clinical commissioning groups. [47]

Clinical commissioning groups (CCGs) are clinically-led statutory NHS bodies responsible for the planning and commissioning of healthcare services for their local area. Members include GPs and other clinicians, e.g. consultants and nurses. CCGs are responsible for approximately 60 per cent of the NHS budget and commission most secondary care services, whilst playing a lesser part in the commissioning of GP services. Secondary care services commissioned by CCGs include planned hospital care, rehabilitative care, urgent and emergency care (including out-of-hours and NHS 111), most community health services, mental health services and learning disability services. [47]

In March 2017, there were 233 recorded NHS providers of secondary and tertiary care, which comprised 152 Foundation Trusts and 81 Aspirant Trusts (additional non-NHS organisations that also provide secondary and tertiary care services). Of the 233 NHS providers, there were 105 acute providers (approximately 45%), providing largely hospital-based services. [46] Additionally, the 'Models of care' programme [48] aims to support improvement and integration of services further, there are five types of vanguard:

Table 2.1 NHS England new care models

Vanguard	**Detail**
Integrated primary & acute care systems	Connecting GP, hospital, community and mental health services to encourage shared goals and incentives to benefit the local population.
Multispecialty community providers	Moving specialist care out of hospitals and into multispecialty community providers.
Enhanced health in care homes	To improve the quality of life, healthcare and health planning for elderly people living in care homes by offering joined-up healthcare and rehabilitation services.
Urgent and emergency care	Approaches to improving the coordination of services and reducing pressure on ED e.g. implementing a new model of integrated urgent care via NHS 111.
Acute care collaborations	Linking local hospitals together to improve their clinical and financial viability, reducing variation in care and efficiency.

Source: NHS England models of care. [48]

The 'Primary care home model' consists of a range of health and social care professionals, including GP surgeries, community, mental health & acute trusts, social care and the voluntary sector, all working together as a complete care community to focus on local population needs and provide care

closer to patients' homes. Thus providing enhanced personalised and preventative care for the local community. [49]

2.2.2 Challenges facing the NHS

The NHS faces many interlinked challenges, including economic, social and political pressures, and an ageing and growing population. The NHS, in effect, treats people, and many diseases which 65 years ago were once fatal have now been cured, which has resulted in increased life expectancy. The combined effect of people living for longer, higher birth rates with lower infant mortality and extrinsic contributory factors such as immigration has resulted in a growing population. [50],

Moreover, an increased life span may inadvertently result in higher chances of people developing one or more long-term complex conditions or illness, e.g. obesity, diabetes, heart or kidney disease or antibiotic resistance. In addition, poor lifestyle choices, e.g. alcohol, smoking, diet, lack of exercise, will also have a negative health impact. As such, healthcare will need to evolve in order to provide ongoing treatment or specialist care for such patients.

Although advancements in medicine and surgery, alongside investment in ICT and technological innovations, help to save lives every year, the rising costs of service provision, energy, supplies and future investment all contribute to unsustainable pressures. The NHS is affected substantially by the economic climate and is extremely sensitive to any financial crisis. As demand for services rise, correspondingly so too do costs. For example, it is estimated that progress in medical technology alone costs the NHS an extra £10 billion per year. Such pressures may lead the NHS to rely increasingly upon privatised services, or the closure of local services in favour of centralisation. [50, 51]

To summarise, the unprecedented slowdown in health service funding, deep cuts to social care services, additional spending pressures (due to rising costs, an ageing population, and an increase in long-term conditions) and rising demand have seen the combined annual deficits of NHS Trusts in England rise significantly. Inadvertently, some hospital trusts have relatively few options other than to offer reduced services due to staff shortage and financial pressures. Given the status quo, one may anticipate potentially detrimental effects on patient care. [45]

2.2.3 Emergency Medicine

Emergency medicine (EM) is the medical specialty dedicated to the diagnosis and treatment of unforeseen illness or injury. Its practice is based on the knowledge and skills required for the initial evaluation, diagnosis, treatment, management of the acute and urgent aspects of illness and injury affecting patients of all age groups, with a full spectrum of undifferentiated physical and behavioural disorders. Disposition of any patient may require expeditious medical, surgical, or psychiatric care and frequently involves the coordination of care among multiple providers. [52] Moreover, it is characterised by the simultaneous care of multiple patients with a wide variety of conditions, a substantial proportion of whom are acutely ill. [53, 54]

EM encompasses planning, oversight, and medical direction for community emergency medical response, medical control, and disaster preparedness. EM may be practised in a variety of settings, e.g. hospital-based emergency departments, walk-in or urgent care clinics, emergency medical response vehicles, at disaster sites, or via telemedicine. [52]

EM traverses an intersection of three major silos: healthcare, public health, and public safety. A silo is “A system, process or department that operates in isolation from others”. [3] EM is a complex nexus of communication and coordination with external services including primary care services, fire and rescue, law enforcement, ambulance services and social services. [55]

2.2.4 Accident & Emergency

Accident and Emergency (A&E) encompasses all types of provision including Type 1, Type 2, Type 3, Type 4 departments and Urgent Care Centres (UCC). Additionally, services that are primarily appointment-based, e.g. GP Practices, Out-Patient Clinics or those that are predominantly accessed via telephone or referral (e.g. most out of hours services) do not fall under the category of Type 3 A&E services even though they may treat a number of patients with minor illness or injury. [56, 57]

A&E departments (EDs) occupy a key position in terms of the interface between primary and secondary care and have a high public profile. [58] An ED attendance is an individual visit by one patient to an ED to receive treatment from the service. [59] Although many patients attend without a referral, nevertheless, some may be referred by other services such as GPs, NHS Direct, Minor injury units, and other medical practitioners. [58]

Table 2.2 Types of Accident and Emergency departments

A&E department	Description
Type 1	Major A&E departments are also known as Casualty or Emergency Departments and are Consultant led 24-hour service with full resuscitation facilities and designated accommodation for the reception of accident and emergency patients.
Type 2	A consultant-led single specialty accident and emergency service (e.g. ophthalmology, dental) with designated accommodation for the reception of patients.
Type 3	An A&E/minor injury department with designated accommodation for the reception of accident and emergency patients. The department may be led by Doctors or Nurse and treats at least minor injuries and illnesses and can be routinely accessed without an appointment. It may be co-located with a major A&E or sited in the community. Other types of departments primarily designed for the receiving of A&E patients include Urgent Care Centres (UCC) and Minor Injury Units (MIUs).
Type 4	NHS Walk In Centres (WiCs).

Source: NHS Data Dictionary [56] and NHS England A&E Attendance/Admissions Definitions. [57]

EDs primarily deal with life-threatening emergencies and manage patients with a huge variety of medical problems, including loss of consciousness, breathing difficulties, acute confused state, seizures/fits, persistent chest pain, severe bleeding, allergic reactions and burns or scalds. Patients frequently present to EDs with painful and/or distressing disorders of recent origin. [58, 60]

Nevertheless, some of the key priorities for EDs are to make life-saving interventions; provide analgesia; identification of relevant issues, undertake investigations, commence treatment and decide upon the need for admission or discharge. [58]

Additionally, patients whose emergency treatment necessitates an overnight stay within the hospital are admitted via the ED of the healthcare provider. Such patient admissions are unpredictable and usually at short notice because of clinical need and are designated as ‘Emergency admissions’ and allocated an NHS code 21. [61] Ultimately, the condition of the patient will determine the medical specialty under which they are to be treated, the treatment plan, and the duration of their stay as an inpatient.

The flow of patients through hospital, from their initial attendance at the ED to eventual discharge, is fundamental to the smooth functioning of an ED, which can be aided by an early senior review of cases. [62] Figure 2.2 (page 31) adapted from Fatovich et al. (2002), [53] visualises the basic processes of patient flow through an ED, beginning with the various avenues by which a patient may enter ED. Once in ED, a high-level overview of the treatment process has been listed, along with the various routes by which a patient may leave the ED.

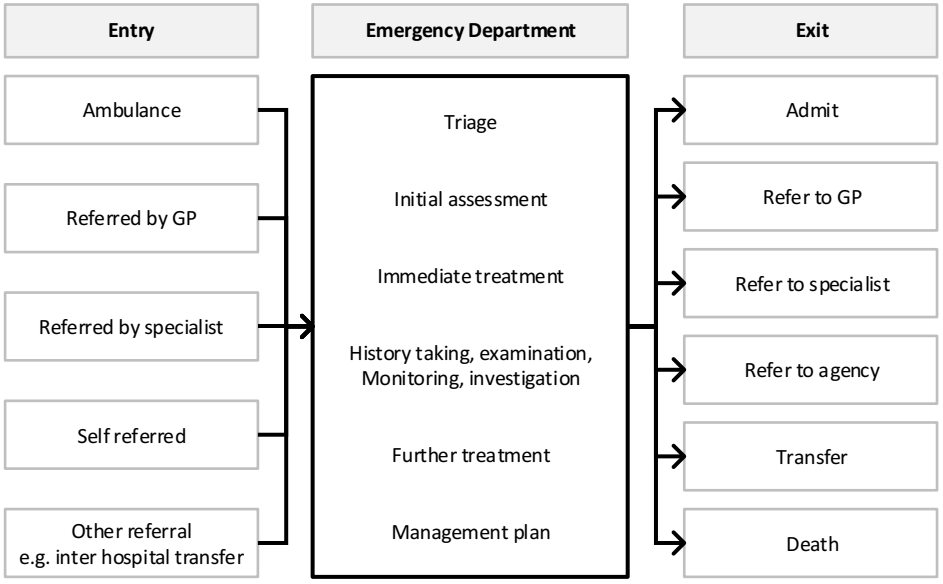


Figure 2.2 Flow of patients in an emergency department (adapted from Fatovich et al. 2002)

2.2.5 A&E Initial Assessment Triage Category

The nature of ED work means that a sorting system is required to ensure that patients with the most immediately life-threatening conditions are seen/treated first; hence, all patients are assessed upon arrival. An initial assessment has three main objectives, the first of which is to improve safety. The second objective is identifying acuity to ensure that the most time-critical patients are treated by the right service within appropriate time frames, and that appropriate prioritisation occurs for the remainder. The third objective aims to improve efficiency in the system to ensure that patients do not wait unnecessarily for investigations or diagnostic decision-making. [58, 63]

The triage process involves an initial assessment by medical or nursing staff and sorts patients, according to a combination of their presenting complaint and measured physiological parameters, at the time of arrival in the ED. The main function of triage is to prioritise patients in a system where the demand for patient care outstrips the ability of the system to deliver it at the time of

presentation. Triage enables a category to be assigned to a patient based on their medical need whilst considering available departmental resources. [63]

The triage category is used to determine the patient's priority for treatment, and to inform the patient of their waiting time. One most commonly used processes in the UK is the National Triage Scale, derivatives of which may also be adopted by individual EDs. [58, 64]

Table 2.3 A&E initial assessment triage category

National Codes		Stream	Description	Colour	See clinician
1	Immediate resuscitation	Resuscitation	Patients in need of immediate treatment for preservation of life.	Red	Immediately
2	Very urgent	Majors	Seriously ill or injured patients whose lives are not in immediate danger.	Orange	Within 5 – 10 minutes
3	Urgent	Majors	Patients with serious problems, but apparently stable condition.	Yellow	Within 1 hour
4	Standard	Minors	Standard A&E cases without immediate danger or distress.	Green	Within 2 hours
5	Non-urgent	Minors	Patients whose conditions are not true accidents or emergencies.	Blue	Within 4 hours

Sources: NHS Data Dictionary [64], NHS Digital [65] and Wyatt 2011. [58]

Upon arrival to ED, patients are assessed by a dedicated triage nurse who in most cases, may take a few minutes, though ideally, triage should be delivered within 15 minutes. The triage nurse may provide immediate intervention (e.g. elevating injured limbs, applying ice packs or splints, and giving analgesia) and initiate investigations (e.g. ordering X-rays), to expedite the patient's journey.

However, triage is a dynamic process, and the urgency (triage category) may change with time if the condition of the patient begins to worsen, thereby necessitating re-categorisation. Placement within a triage category does not imply a diagnosis, nor signify the lethality of a condition.

Resources allocated to the delivery of triage will depend on the number of patients arriving in the ED at any specified time, which inadvertently affects the time allowed to complete the triage process. From a patient perspective, those allocated to non-urgent categories may wait inordinately long periods of time, whilst patients who have presented later with conditions perceived to be more urgent, are seen before them. [58, 63]

2.2.6 A&E Stream

The NHS defines streaming as a classification whereby patients are allocated to different flows according to their needs, e.g. Minors, Majors or Resuscitation. [65] Minors is the term used in ED for cases which are not life-threatening and are usually diagnosed, treated and discharged on the same day, sometimes with follow up. Majors are patients who tend to have more serious, potentially life-threatening conditions, or who require more detailed clinical assessment and investigation by specialist staff. Resuscitation patients require immediate treatment for the preservation of life.

The main objective of streaming is to ensure that patients are directed and allocated to the correct flow, i.e. location/physical area, service, pathway/process so that the correct clinical staff will manage their needs. For streaming to be effective, it is usually performed as soon as possible, ideally within 15 minutes of the patient's arrival to the ED. Hence, it needs to be brief to avoid queues developing. Overall, the streaming process helps improve departmental efficiency and effectiveness.

Each stream is staffed individually and is able to continue functioning almost independently of other streams. At any one time, within any stream, there may be patients allocated to different initial assessment triage categories. There are two overall classifications for the streaming process, and EDs will either employ 'simple streaming' or 'full streaming' classification conventions.

Simple streaming is based on clinical assessment and typically involves the 'streamer' taking a brief history, and performing basic observations and directing the patient to the appropriate physical area of the department, whilst matching patient's needs to departmental capability. It may also be combined with triage, i.e. triage followed by streaming to one of three categories Minors, Majors or Resuscitation.

Full or complex streaming is more advanced and involves a more detailed assessment of priority and acuity as well as ensuring the patient is managed by the correct service within the correct time frame. Additionally, some investigations may be requested to help expedite clinical decision-making processes. [63, 65, 66]

Within full streaming, cases are classified into five categories.

1. Self Care: those patients who attend A&E and do not need investigation or hospital treatment. After a thorough assessment, such patients are given appropriate advice on self-care and are discharged. They may also be advised on indications for further contact.
2. Primary Care: those patients with conditions that could be treated by a primary care team.
3. Ambulatory care, i.e. Minor Injury and Moderate Illness: patients who need some investigation or treatment for an injury but are unlikely to be admitted.
4. Clinical Assessment, i.e. Majors: those patients who require a more thorough clinical assessment and may need a series of investigations or observation before definitive treatment can be determined.
5. Resuscitation: patients with serious or life-threatening illness or injury, which requires emergency assessment and care.

The streaming process may also include streaming to co-located, or specialist services (e.g. co-located primary care services, ophthalmology services) and patients may be streamed (redirected) to services off-site. [63]

2.2.7 Clinical Quality Indicators

The ED Clinical Quality Indicators (CQI), implemented In April 2011, were developed by a collaboration between the Department of Health, clinicians from the College of Emergency Medicine and The Royal College of Nursing. The CQI were designed to provide a measure of the quality of care being delivered in EDs in England. [67] There are eight indicators, grouped into three categories, summarised in Table 2.4 (page 36). [68]

2.2.8 The four-hour standard

Historically, the 'Total time spent in the A&E department' measure, was designed to highlight the importance of monitoring the total time a patient may spend in ED and acted as a driver to help maintain patient flow/throughput. This 'four-hour' standard stipulated that 98 per cent of patients attending ED would spend a total time of less than four hours from the time they arrived up until the time they left the department, via, either being discharged home, admitted to a hospital bed or transferred. [67]

Discharged home is defined as when the patient's clinical episode is finished, and the departure time is when the patient actually leaves the department. A 'Transfer' is defined as transfer to the care of another NHS organisation or other public/private sector agency (e.g. social services). The time of transfer is defined as when the patient leaves the department. An emergency admission via A&E is defined as an A&E attendance against which is recorded either disposal code 1 (the same healthcare provider) or disposal code 7 (transfer to another healthcare provider). The time of admission is defined as the time when a patient leaves the department, e.g. to go to an operating theatre, a bed in a ward or diagnostic test en route to the aforementioned destinations. [57]

The 98 per cent four-hour measure has been superseded by the clinical quality indicators (Table 2.4, page 36), more specifically the CQI which monitors the single longest time, the median (50th percentile) and the 95th percentile of total time spent in A&E from arrival to discharge. Ideally, the 95th percentile should be less than four hours, and any patients with longer times in department essentially 'breach' the threshold, which requires retrospective case review by senior department consultants. Furthermore, to reflect the different processes of care, these measures are reported separately for admitted and non-admitted patients.

> *"The 4-hour target was meant to eliminate the problem of corridors lined with ill and injured patients waiting to be treated by a physician or be admitted to a hospital bed".* [69]

The length of time a patient spends within the ED is dependent upon the type of ED they visit. Minor EDs (Types 3 & 4, e.g. Urgent Care Centres or Minor Injury Units) nearly always treat people in less than four hours. Patients categorised within the Minor stream usually account for 57 per cent of total attendees and approximately 2 per cent of these wait for more than four hours.

Moreover, Major (Type 1) EDs deal with a higher number of attendees and more serious cases. Patients categorised within the Majors stream and not admitted account for 23 per cent of total attendees, of these, 7.5 per cent wait for more than four hours. Major category patients that are admitted to the hospital make up 20 per cent of total attendees, and of these, 23 per cent wait for more than four hours. This inevitably means that Type 1 EDs will find it harder to achieve the four-hour target. [70, 71]

The four-hour waiting standard requires all EDs to see 95% of attending patients within four hours of their arrival. Although, the four-hour standard may intrinsically highlight when pressure in ED is growing and performance suffering, it is not, however, a useful indicator of the quality of care

received by a patient. Essentially, it is a process measure and not a quality measure. Deterioration in performance against this standard has prompted NHS England to investigate initiatives for emergency care. [62]

Table 2.4 Overview of A&E clinical quality indicators

Category/Indicator	Indicator aim
Effectiveness of care	
Ambulatory care	To reduce avoidable hospital admissions by improving the provision of ambulatory care. Calculation: The percentage of A&E attendances for cellulitis and DVT that end in admission.
Unplanned re-attendance rate	To reduce avoidable re-attendances at A&E by improving the care and communication delivered during original attendance. Calculation: Unplanned re-attendance at A&E within seven days of original attendance (include referral back by other health professionals).
Total time spent in A&E department	To improve the timeless & monitoring of care to ensure patients do not have excessive waits in A&E before leaving the department. Calculation: The median, 95th percentile and single longest total time spent by patients in the A&E department, for admitted and non-admitted patients.
Patient experience	
Left without being seen rate	To improve patient experience and reduce the clinical risk to patients who leave A&E before receiving the care they need Calculation: The percentage of people who leave the A&E department without being seen.
Service experience	To improve the experience of patients (& carers) using A&E. Qualitative description of what's been done to assess patient/carer experience of using A&E services. Summarise results and actions undertaken to improve services, in light of these results.
Patient safety	
Time to initial assessment	To reduce the clinical risk associated with the time the patient spends unassessed in A&E Calculation: Time from patient arrival to start of full initial assessment for all patients arriving by emergency ambulance.
Time to treatment	To reduce the clinical risk and discomfort associated with the time the patient spends before their treatment begins in A&E Calculation: Time from arrival to start of definitive treatment from a decision-making clinician.
Consultant sign-off	To improve clinical processes/outcomes and reduce the risk patients are exposed to, for Type 1 and 2 (major) A&E departments. Calculation: The percentage of high-risk patient groups who are reviewed by an emergency medicine consultant before being discharged.

Source: Department of Health data definitions. [68]

2.2.9 National statistics

The Office for National Statistics England population mid-year estimates, [72] indicate that the population is increasing. For example, between the years 2010 (pop 52,642,500) to 2016 (pop 55,268,100) the population had increased by approximately five per cent, which equates to 2,625,600 people. General increases in population are just one of the contributory factors leading to higher A&E attendances. [50, 73, 74] A study by Meggs et al. (1999), [75] looking at trends in ED reported a direct relationship between ED utilisation and population size.

A report by the 'House of Commons' stated that in 2016 there were 23.6 million attendances to England's EDs, of which 65% (15.3 million) were attributed to patient attendance at major (Type 1) departments, the remainder being at minor injury units or walk-in centres. Effectually, total ED attendance had increased by 5.2% when compared with 2015 and appeared to be increasing at a faster rate than population growth. [76]

NHS England [77] retain data concerning monthly A&E attendances and emergency admissions for all A&E types, including:

- Type 1: Major A&E departments
- Type 2: Single Speciality departments
- Type 3: Other A&E or Minor Injury Unit departments.

As this research will focus on Type 1 Major A&E departments, the following results have been restricted to this category.

Table 2.5 Type 1 ED attendances, admissions and four-hour target in England

Type 1 A&E Department	Reference	First value	Final value	Difference
Complete month		08/2010	11/2017	6.5 years approximately
Attendance	Figure 2.3*	1,138,652	1,282,501	143,849 (12.6% increase)
Arrival to gone > 4 hours	Figure 2.4*	2.9%	17.0%	14.10% increase
Admissions	Figure 2.5*	287,438	372,566	85,128 (29.6% increase)

** Figures presented on page 38 of this thesis.*

Table 2.5 and associated figures, provide a brief overview taken over a 6.5-year time frame, leading up to the status quo. Overall, monthly Type 1 A&E attendances have increased by approximately 12.6 per cent (Figure 2.3) with people aged 80+ being the most likely to attend. However, amongst working-age adults, those aged 20-24 are more likely to attend than any other age bracket. [76]

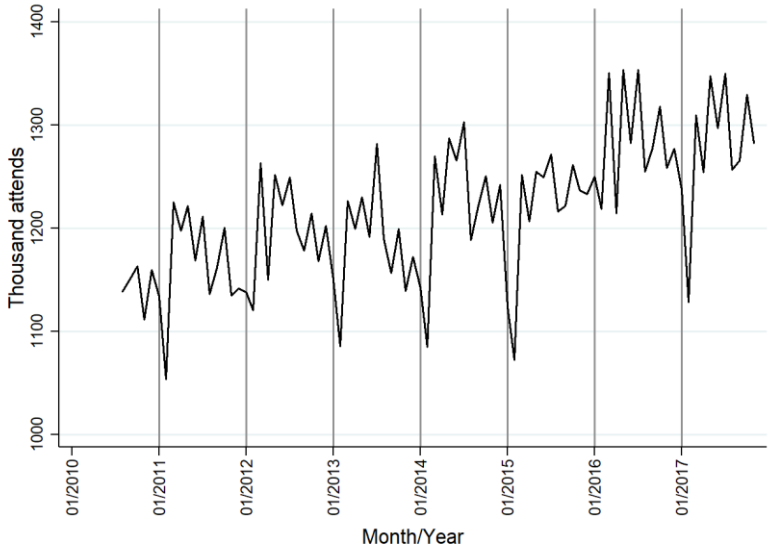


Figure 2.3 Monthly attendance at Type 1 Major A&E departments in England.

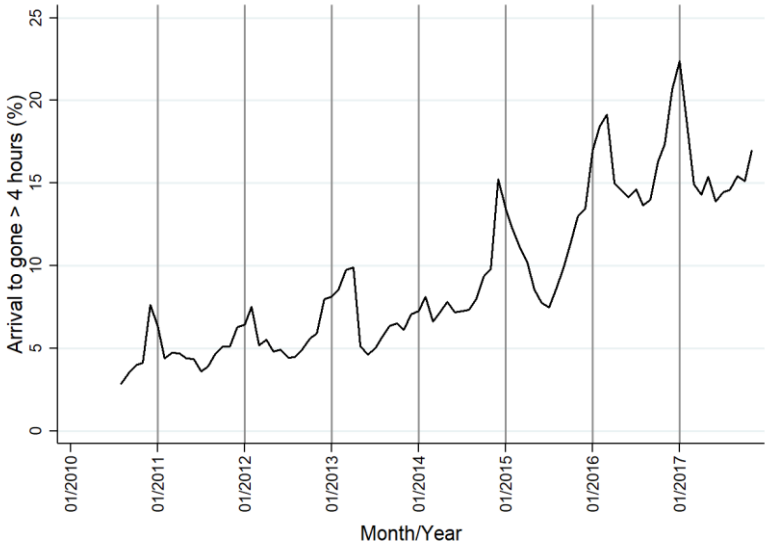


Figure 2.4 Percentage of attendances with a total time in department of greater than four hours, for Type 1 Major A&E departments in England.

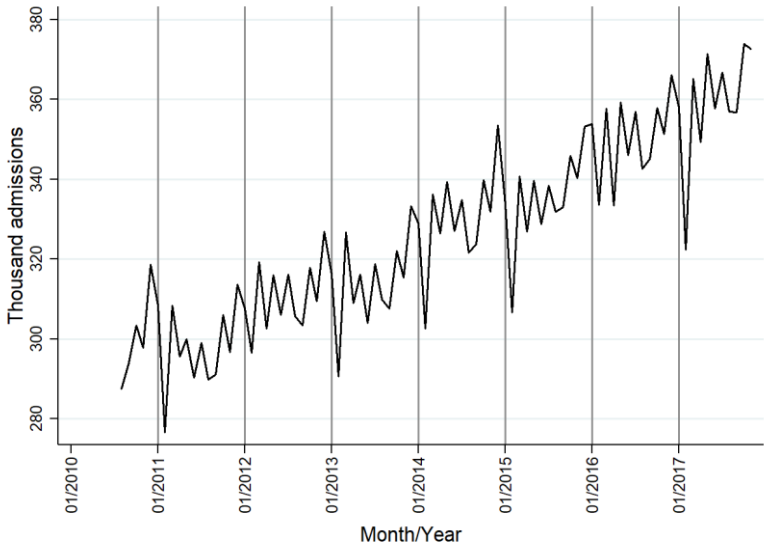


Figure 2.5 Monthly emergency admissions via Type 1 Major A&E departments in England.

Data source: A&E Attendances & Emergency Admission statistics [77]

Correspondingly, the percentage of patients that spend a total time of more than four hours in the department, i.e. Clinical Quality Indicator 3, has seen a five-fold increase from 2.9 per cent to 17 per cent (Figure 2.4), which effectively means that the 95% target is not being met, and stands at approximately 83 per cent. Equally as important is the number of emergency admissions (Figure 2.5), which have increased consistently every year, with an overall increase of almost 30 per cent over the time frame.

Emergency admissions happen at short notice because of perceived clinical need. The National Audit Office [78] report that in 2012-2013 there were 5.3 million emergency admissions to hospitals, which represents around 67 per cent of hospital bed days in England, costing approximately £12.5 billion. Over the last 15 years, short-stay (less than two days) admissions have increased by 124 per cent. Avoiding unnecessary emergency hospital admissions whilst managing those patients that are admitted more effectively is a major challenge for the NHS due to associated costs and pressures affecting elective healthcare. [78]

These figures provide irrefutable evidence that ED attendances are increasing and that this is a growing problem for the NHS, which is not sustainable and requires urgent attention and some form of intervention.

2.2.10 Overcrowding within EDs

Health expenditure in western countries has increased due to a broader provision of health services and progress in the field of medicine. EDs operate daily, all year round, responding to multiple demands that require the utilisation of sophisticated technical equipment and employment of specialist staff, coupled with geographic location and department size, all of which contribute to higher operational costs. [79]

Increasing demand for healthcare far outweighs the limited resources available, a well-known scenario in most ED environments. [80] As life expectancy increases, higher numbers of patients with multiple and complex underlying medical problems present to the ED, ultimately requiring more clinician and nursing time to diagnose and treat. Increased severity of patient illness or injury means decreased turnaround time for beds. [81] Providing care to an increasing number of critically ill patients has contributed significantly to long waiting times and results in overcrowding. [79, 81]

Overcrowding is the most serious issue confronting EDs in the developed world and is an international problem, which is getting worse. Health systems in the United Kingdom, Canada, United States, Australia and New Zealand are all facing similar difficulties, which suggests that a profound change has occurred in the delivery of Western medicine. [53]

In simple terms, if patient inflow to ED is greater than patient outflow, the result is overcrowding. EDs may experience an overwhelming number of patients presenting in a short period of time, whilst dealing with a lack of beds for acute patients. [53] Overcrowding in EDs is difficult to define scientifically, some emergency physicians define overcrowding as a situation in which demand for service exceeds the ability to provide care within a reasonable time, causing physicians and nurses to feel too rushed to provide quality care. [82]

Consider the analogy of a glass filled with water; if additional water is poured into the glass, it simply overflows. [81] Overcrowding in ED is the biggest impediment to the delivery of timely and adequate emergency care and is caused by a complex web of inter-related issues.

2.2.10.1 Causes of ED overcrowding

The overall complexity and acuity of patients presenting to EDs have increased, and a large proportion of the population is ageing, medical technology and better pharmaceuticals have enabled people with chronic diseases to survive longer, e.g. patients with congestive heart failure, AIDS or renal failure. However, these conditions may become exacerbated and necessitate patients to seek ED care. Additionally, patients presenting with several comorbid diseases, complicate the evaluation process and require more time for appropriate investigation and subsequent treatment. [81, 82]

Many EDs have experienced an overall increase in patient volume, which could be attributed to regional population growth. Consequently, patients with no place to go for medical care frequently attend ED for urgent and primary care needs, e.g. patients not registered with a GP. Additionally, managed long-term care patients may face difficulty when booking appointments or accessing their personal care clinicians due to limited immediate availability. [81, 82]

Another very important cause of ED congestion and overcrowding is a lack of beds for patients admitted to the hospital. Bed shortages commonly involve an inadequate number of ICU (Intensive Care Unit) or telemetry beds. Furthermore, if medical-surgical beds become completely filled, ED

patients that require hospitalisation must wait until a bed becomes available, thus occupying valuable space in the ED. [53, 82]

There is an expectation that ED physicians are able to provide observation areas and offer extended treatments in order to stabilise and discharge patients in lieu of hospital admission. The administering of 'intensive therapy' within the ED may help to avoid inpatient hospital admissions. Advances in medical science, improvements in technology, new pharmacologic agents, heightened standards of practice, and residency training of emergency clinicians have all contributed to this expectation. [73, 82]

As a consequence of advancing technology and escalating standards of care, more patients in the ED need plain radiographs, laboratory tests, or other types of specialised ancillary services, e.g. CT scans, magnetic resonance imaging (MRI) or arterial blood gas values. Depending on the size of the hospitals and the number of patients, such services will also be affected, thus compounding total patient time in the ED and leading to delays in service provision. [53, 82]

Another contributory factor is the lack of availability or shortage of on-call specialty consultants. Consultations by specialists are necessary to provide definitive treatment in the ED or to admit patients and a shortage of such specialists will inevitably lead to a delay in patient care. In some cases, the overwhelming amount of inpatient and clinic responsibilities of the on-call specialist may lead to serious delays in responding to the ED. Furthermore, in the case of smaller community hospitals, certain on-call specialties may not be available, and ED patients must be transferred to a larger hospital, which effectually causes even further delays. [81, 82]

Similarly, experienced and dedicated nursing staff are critical to optimal functioning and care provision within an ED. A shortage in nursing staff will mean inadequate numbers remain to provide for increasing numbers of patients. Another interrelated problem is high staff turnover, which effectively leads to a higher proportion of new, inexperienced emergency nurses. [73, 81, 82]

Many ED functions rely on administrative/clerical support staff to undertake patient registrations, telecommunications, process paperwork, make requisitions for laboratory, radiology and admissions. Reductions in the number of clerical support staff will negatively affect patient flow. [81, 82]

The physical size and layout of EDs are limited, leading to a restricted amount of space available for patients. The average length of stay for patients presenting to the ED is increasing because of a lack of beds; hence, a larger ED would be required to facilitate the same number of patients. This problem is further compounded when the volume of ED visits increase. [73, 82]

EDs in geographic areas with a diverse ethnic population or a high proportion of immigrants may evaluate patients who have little or no understanding of the English language and therefore require additional time, e.g. to obtain a translator. Even when a translator is available, the interaction between patient, translator and emergency healthcare staff will require more time. Depending on the severity of the problem, such language and cultural barriers may complicate and prolong diagnosis and treatment, which again contributes to an overall decrease in patient flow. [82]

Furthermore, increased medical record documentation requirements, e.g. for medico-legal purposes, mean that emergency healthcare staff may spend more time with the 'chart' than with a patient. Similarly, it may be difficult to arrange follow-up care for patients not registered with a GP. [53, 82]

2.2.10.2 Effects of ED overcrowding

Nonetheless, there are many effects of overcrowding within the ED and the poor outcomes that result from such conditions in some hospital EDs put public safety at risk. Healthcare staff treat more complex and acutely ill patients and increasingly perform non-clinical tasks. Hence, they may feel under time pressure and rushed, which may not give them adequate time to thoroughly evaluate each patient, ultimately resulting in errors, which risks poor patient outcomes, malpractice or even legal action. [53, 82]

Overcrowding also has a negative effect on physicians and leads to decreased productivity. Although emergency physicians attempt to fill in the gaps, they must stretch their ability and aim to see many patients at the same time. However, at a certain limit of patients, productivity will decline, and patient care will inevitably be compromised. Medical staff may feel an increasing sense of frustration, and as a result, might decide to seek employment elsewhere. [53, 82]

Due to periods of overwhelming patient volume, faulty communication amongst healthcare staff may occur, which unavoidably results in decision errors, e.g. mislabelled specimens or radiology request forms. Medical teaching within academic medical centres involves a follow and watch

approach with more formalised bedside teaching. However, the latter has been negatively affected; as in busy EDs, senior consultants may be too busy with patient care to teach medical students. [73, 82]

During times of overcrowding, patients face longer than normal waiting times, which leads to them experiencing prolonged pain and suffering unnecessarily because ED staff are too busy to attend them. These problems inevitably lead to patient dissatisfaction, which manifests itself as the patients who leave ED without being seen or treated. Consequently, this has the potential for minor medical problems to become more serious from delay in care. Additionally, prolonged waiting times can cause patients to become more agitated and violent, as tempers flare, patients may fight amongst themselves about who was first to be seen or may even threaten/harm healthcare staff. [53, 82]

From a safety perspective, there may be an inability to evacuate patients or staff safely in an emergency, such as a fire. Within urban areas, EDs may instruct the ambulance service to divert ambulances elsewhere because it is unsafe for more patients to attend, which results in significantly increased transport times, risk of traffic accidents en-route, and increases the potential for poor clinical outcome. [81, 82]

To summarise, for people that require non-elective acute hospital care, EDs are the main point of entry into inpatient settings. [53] Taken together, the many causes and effects due to overcrowding in EDs have negatively affected physician productivity and efficiency. [82] Previous research by Derlet et al. (2000), [82] highlights the importance of having real-time computerised tracking of waiting times, treatment times and current census of the actual number of patients in the ED being treated or waiting to be seen, as such information may help to define overcrowding accurately.

2.2.11 Emergency admissions

There are many interlinked and multifactorial factors, which contribute to increases in emergency admissions. The four-hour standard was initially introduced by the Department of Health to enable ED to focus resources, improve decision-making processes and reduce waiting times. However, it may have inadvertently contributed to the increase in short-stay emergency admissions, which is because the four-hour standard reduces a hospitals' ability to keep a patient in the ED for monitoring and observation, unintentionally influencing the admission behaviour of EDs.

Almost a quarter of admissions via ED occur within the last ten minutes of the four hour period, which may also reflect a hospital's mechanism for managing limited bed capacity by delaying admissions until close to the four-hour mark. [78]

Changing medical practices and models of care is another important contributory factor. Some elective procedures have seen a shift from inpatient care (where a patient has a procedure and then stays in hospital for one or more nights), to, day-case care (where the patient undergoes the procedure and goes home the same day), a minority (approximately 3 per cent) develop complications leading to emergency admission.

Therefore, the increase in day-case activity has led to a corresponding rise in emergency readmissions due to complications. The National Audit Office analysis of hospital episode statistics identifies that these readmissions can explain approximately 9.4 per cent of the increase in emergency admissions. [78]

Other reasons include increased uses of clinical protocols and lowering of clinical thresholds, which may lead to the admittance of less severe cases, and the increased threat of litigation may possibly lead to defensive medicine. [78] Nonetheless, the constant assessment and reassessment of patients by junior staff in medical assessment units only breeds duplication, which delays authoritative decisions regarding treatment, transfer or discharge. [62]

An additional factor is a change in the patient demographic, which includes an increasingly frail elderly population living with one or multiple long-term conditions. The elderly are far more likely to have immediate or chronic health problems requiring urgent care, attend ED and require subsequent admission. Urgent access to primary care is variable, and this may result in higher ED attendances. As such, the tendency to admit patients is greater when EDs face increased pressure. Furthermore, the severity of cases presenting to ED has worsened, with higher proportions of patients arriving via ambulance requiring admission. [78]

Various strategies designed to help reduce emergency admissions include:

- Better management of a patient's long-term condition by primary, community and social care services.
- The conveyance of patients to a wider range of care destinations by ambulance services.

- Hospitals ensuring prompt initial senior clinical assessment, quick access to diagnostics and specialist medical opinion.
- Following admission, hospitals can work with community and social care services to ensure patients are discharged promptly, staying no longer than necessary.

Systemic issues arise due to a lack of comprehensive alternatives to emergency hospital admission; hence EDs are perceived as the default route for urgent and emergency care. [78] Nevertheless, investigating the rise in emergency admissions requires a multifactorial model; however, it is yet unclear as to what particular admission rate is deemed acceptable. Thus, any interventions will require assessment and proper evaluation to determine the extent of their benefit. [83]

2.2.12 Patient preference for ED

In 2016, there were 23.6 million attendances to EDs within England, and of these, 15.3 million (65%) were attributed to patient attendance at Type 1 Major A&E departments. [76] Included within these attendances are a significant number of patients who refer themselves to ED for conditions which are neither accidents nor emergencies, and relatively few of these patients require specific hospital treatment.

Patients with conditions that do not befit A&E have been described as ‘inappropriate’, however, there is no accepted practical definition of what constitutes an ‘appropriate’ attender to ED. [84] One fundamental limitation within previous research about ‘inappropriate’ attendees revealed that such patients could only be labelled through retrospective comparison with ‘legitimate’ patients.

This was accomplished by looking at the final diagnosis, which helped identify those problems/conditions that did not warrant A&E services [85] and could have been managed within primary care. [86] As such, patients who show a strong preference for ED, mainly based on assumptions of quality of care can be collectively termed ‘self-referrals’. [85]

Self-referrals perceive that EDs are the most appropriate place for their problem and that hospital specialists are best qualified to handle their problem. [85] They want a specialist opinion and ‘a thorough check-up’ with lengthy physical examinations. Additionally, self-referrals anticipate undergoing a battery of investigations as they believe that EDs possess the necessary diagnostic

facilities and advanced technology, which they perceived the quickest way to obtain was via ED. [85, 86]

Furthermore, the process of decision-making for patients opting to attend ED in preference to consulting their GP is complex and involves an interplay of social, psychological and medical factors such as the appropriateness of the condition, anticipation of referral, GP availability, ED accessibility, and in some cases, advice given by another person. [84]

One of the most important factors that influence the decision of a patient to attend ED is perceived severity. This refers to the patients' perception of the urgency of their illness, which stems from both personal beliefs and knowledge of what is an emergency. Patients who feel their condition is urgent have a tendency to seek immediate treatment. [87]

A previous study by Uscher-Pines et al. (2013), [87] reported age as a critical factor associated with non-urgent ED use, as younger adults were more likely to have non-urgent visits when compared to older adults. Additionally, the ethnicity of a patient also played a part, and the study indicated that people of black ethnic origin were more likely to have a non-urgent visit, even though they were less likely to utilise healthcare in general, compared to people of white ethnic background. [87]

Additional factors that may influence the decision of a person to attend ED include patient gender, personality, individual emotional/attitudinal and behavioural response patterns along with cultural/community norms, e.g. tendency for friends or family to visit ED. Furthermore, persons with poor health and/or low incomes were found to be more likely to make non-urgent ED visits, whilst some patients may even harbour negative perceptions of non-ED care. [84, 86, 87]

Convenience, in this context, refers to the ease with which a patient can seek care and this includes service availability, i.e. EDs being open and accessible 24 hours/day all year round when compared to other ambulatory care alternatives. For patients that work or have set commitments, timing is very important and such patients may not want to wait for a GP appointment, nor disturb the GP out of hours, and thus prefer visiting an ED so it can fit with their schedule. [86, 87]

Self-referrals want to be seen at a time that is convenient for themselves, i.e. during the evening or at weekends, which helps them to avoid taking time off work. Furthermore, some self-referrals may

anticipate seeing a specialist on the same day for an acute problem or be placed on an outpatient waiting list for a chronic problem. [84, 86, 87]

Proximity to the ED, patient residence, geographic location and travel, are factors that the patient will also consider. For example, if a patient happens to live geographically closer to an ED when compared to the proximity of their GP, or perhaps if the patient happens to pass by an ED to/from their place of work, the patient may deem ED attendance more convenient, which subsequently leads to higher use of this service. This effect may be more pronounced in rural and suburbanised areas than in larger cities. [84, 85, 87]

Access, in this context, refers to the ability of a patient to obtain timely care outside the ED. From a non-urgent ED patient perspective, access may be considered to be poor if the patient has difficulty when attempting to obtain regular healthcare. For example, patients' perceived urgency versus the availability of GP appointments, which is the most significant reason for a patient seeking emergency care. Additionally, patients that are not registered with a regular physician/GP practice or those with limited access to physicians on weekends tend to make more use of EDs. Whereas, in some cases, patients may be counselled, advised or even referred to the ED by another healthcare provider. [84-87]

Nonetheless, previous healthcare service experiences of a patient, both within and outside of the ED, will ultimately determine their future utilisation behaviour. If a patient previously had a good experience in the ED, then they may be inclined to attend more frequently, which ties in with personal beliefs and knowledge about alternative healthcare availability. Additionally, patients who are not satisfied with their regular source of care will be more inclined to make a non-urgent visit to an ED. [85, 87]

Although educating patients to make more appropriate use of NHS services is necessary, it may not result in an effective change in the self-referral rate to EDs as it addresses the behaviour of a person when taken in a rational context, rather than in circumstances where the person is affected by pain and anxiety. [85]

Furthermore, certain groups of patients, e.g. homeless, elderly, children, drug addicts, patients with mental illness tend to use ED disproportionately, and many older people end up in EDs without any genuine clinical need for this type of treatment. [62, 79] Thus, it is the patient, who through

perception, assessment and urgency of their condition, who ultimately decides where to seek appropriate treatment. A better understanding is therefore required, of patients' incentives to attend EDs and the extent to which their demand is complementary to the demand for GP care. [85]

2.2.13 Government urgent care service initiatives

Throughout the NHS, primary care clinicians have the most interaction with patients, and a significant proportion of ED work could more appropriately be classified as primary care and undertaken by GPs. [88] However, it is unclear if primary care has the capacity to absorb additional work.

Walk-in-Centres are able to cater for demand outside of ED and traditional GP surgeries, however, inconsistent opening hours and lack of facilities, e.g. Radiology services, ultimately mean that they are sub-optimally used. Attempts to provide better access to primary care services via the development of walk-in-centres have resulted in an overall increase in demand, induced by the provision of additional services. Theoretically, urgent care can be delivered within a primary care setting, but this would require substantial restructuring of existing services. Thus, as there are no clear solutions for each local health system, clinicians are best placed to understand what works well and what can be replicated elsewhere. [62]

NHS 111 is a non-emergency telephone service based around triage by call-handlers who are not clinically trained. Call handlers use the NHS Pathways IT system to assess patient symptoms and can refer patients to a nurse, emergency dentist or general practitioner if required. The service launched in England in 2013, without any real understanding of the impact it would have on other parts of the NHS, including emergency and urgent care. [62] One criticism is that the service is excessively risk-averse and that triage may be so off-putting to patients that they prefer the option of going directly to ED. In its current configuration, NHS 111 does not help to prevent people from inappropriately attending ED. However, looking forwards, the NHS is aiming to raise the proportion of calls that receive a clinical assessment to approximately 30 per cent. [89]

Thus, to enhance the overall system of emergency care in England, ambulance services should be regarded as a care provider and not a service that simply readies patients for journeys to the hospital. Skilled paramedics can play a significant role in providing emergency and urgent care. For example, treating patients at the scene to reduce conveyance rates to EDs and making difficult judgements about when to by-pass the nearest ED in favour of specialist units. Effectively, this would

contribute to alleviating some of the pressures in EDs, offering a better service to patients, particularly in rural areas, where journey times are longer and need consideration. Moreover, increasing the number of fully qualified paramedics may help to achieve this objective. Ultimately, ambulance services have the potential to coordinate other elements of the emergency and urgent care system and lead integration of services; however, like EDs, they too are facing ever-increasing demand. [62]

A prospective approach to instilling efficiency and providing some clarity in the provision of urgent and emergency care services may be to co-locate Urgent Care Centres (UCC) with EDs on hospital sites, where geographically practical. In this respect, UCCs co-located with EDs provide an opportunity to stream patients with less serious illnesses and injuries to a service that is resourced to meet their needs, while reducing crowding in EDs. [88] Thus ensuring that patients are seen/treated by the most appropriate service. However, to ensure proper service integration shared governance arrangements, and clearly defined protocols for the two-way transfer of patients will need to be established.

2.2.13.1 What is causing pressures in urgent and emergency care?

The successful provision of emergency and urgent care is a matter of life and death, as increasing numbers of patients with more complex and acute conditions present to A&E. Similarly, the number of emergency admissions continue to rise which cannot be simply explained by ambulance services conveying too many patients, nor the NHS 111 service directing too many patients to A&E, nor the reduction in bed capacity, which limits hospital flexibility. As such, there is no sufficient explanation as to why emergency care is operating under such sustained stress. [62, 70]

Additionally, emergency care service provision is, by nature, fragmented, which has inevitably resulted in a blurring of the lines of responsibility and accountability for funding and managing such services. Likewise, patients seeking treatment may potentially become confused by which health service is most appropriate for their needs and just opt to attend A&E by default.

Investigations into understanding the root cause as to why urgent and emergency care services are faced with such substantial pressures have resulted in the identification of not one, but many different facets of the health system contributing to the overall problem. Hence, the problems cannot be attributed to any one specific factor or failure within the health system. [62]

Consequently, the current system cannot accurately analyse the cause of the problem, let alone attempt to resolve it, and as such, it is 'flying blind'. [62] Evidence regarding the profile of patients presenting at A&E is contradictory, and there is a pressing need for clearer intelligence, which can detail where cases present across the system and the case-mix of such presentations. This will provide a sense of direction for understanding trends in the level and nature of demand for urgent and emergency care, and to achieve this, information is required. The use of accurate and clearer information will help provide detail as to where urgent care cases present across the system and the case-mix of urgent patient presentations. [62, 70]

Moreover, accurate information is of critical importance to help monitor waiting times for urgent and emergency services in order to ensure that services are made available and accessible to patients in urgent need of care. Accurate information about the causes of rising service pressures is not simply a management convenience; it is fundamental to the delivery of high-quality care. [62]

2.2.14 Primary Care in ED

Internationally, the inappropriate use of ED is a problem [90] and making international comparisons is intrinsically difficult due to disparities in healthcare systems, facilities and even in the definition of 'inappropriate' attenders. [37] Thus, healthcare providers and patients have differing views about the appropriateness of ED use. [90] Nonetheless, health services have attempted to reduce the number of self-referrals or inappropriate ED attenders by encouraging patients to seek other sources of care before presenting to EDs, e.g. the availability of walk-in centres; better patient education; use of triage services to identify and redirect inappropriate ED patients to the most appropriate service, including referring back to their GP.

Employing a screening or referral process may help to reduce the number of patients that re-attend and the overall length of stay, along with subsequent admissions to hospital. However, such attempts do not appear to have been particularly successful. [79, 85, 91]

It is widely accepted that the use of ED for minor health problems is more expensive than a consultation with a traditional family physician GP. [79] The redirection of non-urgent visits to alternate settings is a desirable policy goal, if for no other reasons than to reduce healthcare spending and enable EDs to focus their efforts on more acutely ill and injured patients. [87] Sessional

GPs working in ED manage non-emergency attenders safely and use fewer resources than do usual ED staff. [91]

Reports looking at primary care clinicians within EDs, published by the Primary Care Foundation [66] and NHS Interim Management and Support Team [92] respectively, identified four main operational models that have been summarised henceforth.

2.2.14.1 Model one: GPs within the ED.

Patients attending the ED and are triaged into separate streams (urgent or non-urgent/primary care). The primary care stream is staffed by primary care clinicians, which helps to supplement shortages of hospital staff. This model provides patients with the most appropriate level of service and enables the main ED to focus on higher acuity patients, which may help to reduce waiting times for all streams and increase patient satisfaction. [92]

2.2.14.2 Model two: The location of GP services alongside the ED.

Primary care services are distinctly available next/close to the ED, and patients either choose themselves or are redirected from the ED towards the primary care service. This model may help to reduce overcrowding in main EDs and has similar advantages to Model one. [92]

2.2.14.3 Model three: Primary care front-end screening/filtering ED patients.

Primary care practitioners are involved in the triage or filtering of non-emergency patients at the front of the ED, by screening, treating or diverting attendees appropriately, thus preventing them from accessing the ED. This ensures that only patients assessed as requiring emergency treatment will be referred to ED, thus reducing overcrowding whilst simultaneously permitting ED clinicians to focus on higher acuity patients. This delineation of services provides clarity of responsibility between the services. [92]

2.2.14.4 Model four: Primary care services fully integrated with ED.

The complete range of care is provided jointly with ED staff to all groups of patients who attend the hospital, including higher acuity and primary care patients. This is well suited for urban areas with highly mobile populations or hard-to-reach groups, i.e. drug/alcohol users, patients with mental health issues, children, frail and patients with long term conditions. [92]

The requirements for primary care services in EDs will vary widely across the country, e.g. urban or rural areas. However, such services should be designed to address specific local needs.

Table 2.6 Examples of primary care involvement in ED

Organisation	Detail
Imperial College Healthcare NHS Trust	Incorporates model two whereby primary care services are located alongside ED, and Model three, with GP led UCC services located at the front-end of ED.
Kettering General Hospital NHS Trust	Integrated urgent and emergency care centre, primary care clinicians manage the majority of paediatric attendances and most patients arriving from care homes.
Nuffield Health Hospitals Chichester & Exeter	Several EDs have GPs as part of the team seeing 'majors' patients. Their local knowledge enables patients to be managed without an overnight hospital stay and creates learning opportunities for junior doctors.
Whipps Cross University Hospital, Barts Health NHS Trust	Has a long-established pathway where appropriate patients are streamed to the primary care out of hours service 24-hours a day.
Sunderland Royal Hospital	A nurse navigator directs appropriate patients to the on-site urgent care centre for treatment led by emergency care practitioners.
Blackpool Teaching Hospitals NHS Foundation Trust	Pathways have been used to support the navigation of walk-in patients to the right emergency care stream, including the on-site urgent care centre.

Source: NHS Primary Care in EDs [92]

Previous research looking at self-referrals to ED in the Netherlands by Moll van Charante et al. (2008), [85] reported that health policy-makers and insurance companies were propagating the integration of GP cooperatives and EDs into one facility. The authors maintained that such a close collaboration had become necessary, as patients with urgent out-of-hours problems felt indecisive about whom they should contact. Furthermore, this new model would help prevent patients from bypassing the GP cooperative and self-referring to EDs as most patients presented with problems or conditions that could be treated equally well by GP services, thereby avoiding unnecessary crowding in ED. Thus, integration of services would lead to more efficient use of resources at a lower overall cost.

Nevertheless, new integrated care models that can adequately deflect those who are eligible for GP care require further research. [85] Another experimental study looking at the effectiveness of GPs in ED within the Netherlands, by Boeke et al. (2010), [90] reported the majority of self-attenders were treated adequately by a GP, with no decrease in the quality of diagnosis. The authors also reported a

positive effect had been observed on patient satisfaction, time management, efficiency, and cost-effectiveness of care.

In contrast, some hospitals have employed GPs within ED to take care of patients that present with 'primary care' type conditions, and UK based studies have shown that GPs working as an integral part of the EDs can work equally safe with fewer resources and at a lower cost. [85]

2.2.14.5 Study 1. Cost-effectiveness of treating primary care patients in A&E

Dale et al. (1996), [93] undertook one of the first economic analysis studies in Britain comparing outcomes and costs of General Practitioners (GPs), Senior House Officers (SHOs), and Registrars treating patients who attended EDs with problems that were assessed at triage as being of a primary care type. The study took place within the A&E department of King's College Hospital, which is an inner-city hospital situated in southeast London.

This prospective intervention study was undertaken for 12 months, through the hours of 10 a.m. to 9 p.m. and included 4641 patients (41%, N=11320) all classified as presenting with primary care problems. Of the 4641 patients, 37 per cent (n=1702) were seen by GPs, 51 per cent (n=2382) by SHOs and 12 per cent (n=557) by Registrars.

Hospital costs were retrospectively allocated and the average (mean) cost per case was derived, by taking into account various categories including Staff time (consultation and transactions); Diagnostic tests (e.g. radiology, microbiology); Treatment and referrals (e.g. hospital admission, outpatient). The average costs per case were calculated for SHOs, Registrars and GPs as £19.30, £17.97, and £11.70 respectively (excluding the cost of admissions) and £58.25, £44.68, and £32.30 respectively (including the cost of admissions).

The key messages stated were, improving the cost-effectiveness of treating patients attending A&E departments was important to both purchasers and providers of services. The comparison of costs and outcomes, of GPs and hospital doctors treating patients attending A&E with primary care problems revealed that there were no significant differences between the types of 'doctor' in terms of clinical outcome and patients' satisfaction. GPs were able to provide care more cheaply than hospital doctors, which was reflected by less frequent requests for investigations and referrals. Thus, employing GPs in A&E departments offers a potential means of reducing cost.

The study concluded that the management of patients with primary care needs in A&E departments by GPs reduced costs with no apparent detrimental effect on the outcome. This result supports a new role for GPs. Furthermore, during a consultation, GPs were able to advise patients about the importance of continuity of care and encouraged patients to attend their own general practice for review and further care. [93]

2.2.14.6 Study 2. Prospective identification of patients

Another associated study by Dale et al. (1995), [94] focused on the triage criteria for primary care and A&E attenders. The authors categorised patients into two classifications:

- Primary care attenders: were defined, firstly as self-referred patients with symptoms likely to be caused by conditions that do not necessitate immediate resuscitation or urgent care and unlikely to require hospital admission, and secondly as self-referred patients with non-urgent complications of chronic conditions.

- A&E attenders: were defined, firstly as all patients referred by letter or phone by a GP, secondly as all emergency presentations in need of immediate care or likely to require hospital admission and lastly as trauma requiring urgent hospital assessment (e.g. fractured bones, dislocations, head injuries with loss of consciousness).

There were found to be several differences in patient characteristics and consultation activities between the two classifications. Presentations made by primary care attenders were less severe, tending to be problems of longer duration, and mostly not related to an injury. In contrast, patients classified as A&E attenders were more likely to be referred to on-call teams; received almost twice the number of radiographic investigations, and had more than three times the number of haematology and chemical pathology tests when compared to the primary care attender classification.

The study identified that A&E triage could be developed in order to identify patients with problems more likely to be of a primary care type. Such patients were less likely to receive an investigation, minor surgical procedure, or referral. However, many of these patients did receive interventions which most likely supported their decision to attend A&E rather than general practice. [94]

2.2.14.7 Study 3. Comparison of general practitioners and hospital doctors

A further related study by Dale et al. (1995), [95] compared the process and outcome of primary care consultations undertaken by SHOs, Registrars, and GPs in an A&E department within the same study setting [93] but for time frame between 01st June 1989 and 31st May 1990.

This study recognised that considerable differences existed between the consultation practices of GPs, SHOs and Registrars in the provision of primary care in an A&E department. Primary care consultations made by SHOs and Registrars involved considerably greater utilisation of investigative, e.g. laboratory and radiography, outpatient and specialist services/resources, than those made by GPs. Therefore, employing GPs in A&E departments to manage patients with primary care needs can result in reduced rates of investigations, prescriptions, and referrals. The results of this study suggest important resource implications which arise from the employment of GPs within A&E.

Thus, like general practice, A&E departments have a gatekeeping role at the interface between hospital and community-based services. However, the stressful and often hectic environment within A&E may make the provision of quality primary care challenging to achieve. Nevertheless, achieving more community-oriented provision should enable greater integration of the care provided by both hospital and community-based health services. [95]

One approach to GPs and A&E departments collaborating involves the co-location of GP services next to A&E departments. Whereby, each service maintains its own professional identity while facilitating efficient referral between the two, thus, reinforcing the gatekeeping role and position of the GP. [85] The use of ED for non-urgent conditions inevitably leads to excessive healthcare spending, unnecessary testing and treatment, and weaker patient-primary care provider relationships.

GPs working as an integral part of an A&E department can manage non-emergency A&E attenders safely, whilst using fewer resources compared to usual A&E staff. [93, 96] For health policy-makers and providers wanting to reduce the use of the ED for non-urgent problems, interventions must be sought that are evidence-based and tailored to address the needs and concerns of the populations they are designed to serve. [87]

2.2.15 Section summary

This section provided the background to emergency healthcare services within the NHS in England and described the challenges faced concerning A&E service provision, Government standards and initiatives such as the four-hour target. The issue of increasing trends in emergency attendances and admissions faced by Type 1 Major EDs within England was highlighted.

Furthermore, attention was drawn to other issues facing EDs, such as patient preference for ED and overcrowding brought about by patients presenting with primary care type conditions. Finally, the concept of four new models, whereby primary care GPs worked alongside EDs to alleviate pressures, were presented and evidenced by previous research undertaken by Dale et al. (1995 and 1996). [93, 95] The authors reported that GPs located within EDs were able to manage non-emergency A&E attendees safely, and use fewer resources when compared to usual A&E staff.

2.3 Data fragmentation

2.3.1 Health information fragmentation

Information can be defined as 'facts provided or learned about something or someone' and 'what is conveyed or represented by a particular arrangement or sequence of things'. [97] This simplistic definition emphasises the importance of arranging facts into a sequence in order to make them coherent.

Moreover, information is an ethereal commodity and from a clinical perspective can be thought of as organised patient data and clinical knowledge that intelligent systems (human and artificial) use to reduce uncertainty, support clinical decisions and guide actions (adapted from Wyatt, 2005 and Shortliffe, 2014). [98, 99] NHS England describe information as its lifeblood, a vital resource for the clinical management of individual patients and the efficient management of services. [25]

What exactly constitutes information? The author Ackoff (1989), [100] identifies five paradigms: Data, Information, Knowledge, Understanding and Wisdom. The first four paradigms relate to the past, i.e. they deal with what has been or what was known, only the fifth paradigm wisdom, deals with the future as it incorporates vision and design.

Commonly in speech, the terms data, information and knowledge are often used interchangeably. However, each term has a quite precise and distinct definition within the field of information science. [101] The difference between data and information is functional, not structural. Data (plural of datum) are essentially symbols representing the properties of objects and events, i.e. data are observations or measurements about the world, for example, 'today is Friday', the 'blood pressure is 125/70 mm Hg' or 'this drug is penicillin'. [100-102]

Whereas information is processed data that has been given meaning via a relational connection, i.e. it is the meaning obtained by the application of knowledge to data, e.g. the datum 'blood pressure is 125/70 mm Hg' provides information only if it tells us something new. Information is contained within descriptions and seeks to provide answers to questions such as who, what, where and when. [100-102]

Knowledge is the appropriate collection and application of information conveyed via instructions, seeking to answer 'how-to' type questions. Understanding is conveyed by explanations, providing answers to 'why' questions, also the synthesis of new knowledge using existing knowledge. By using a piece of knowledge in a given context, data can be interpreted to produce information. Knowledge defines relationships between data, for example, the statement: 'penicillin is an antibiotic', relates two data elements. [100, 102]

Wisdom is the ability to increase effectiveness and involves the exercise of judgment. [100, 102] Decisions inform our actions, and to make sensible/logical decisions information that facilitates the differentiation of one course of action over another is ultimately required. [101]

Thus, patient-specific knowledge comes from observing a patient's state over time. Clinicians take data in the form of patients' observations and measurements and use knowledge about disease and therapy to infer what illness a patient may have and decide on the most prudent course of action. [101]

Healthcare organisations are information-intensive enterprises, and clinical personnel require data and information to make appropriate decisions to provide patient care. [6] Clinical information concerning the nature of the disease or symptom, the treatment decision and outcome is commonly recorded very meticulously.

Moreover, information can take many forms; a clinical datum may be a single observation of a patient, e.g. temperature reading, a red blood cell count, a history of rubella or a blood pressure reading. Whereas, clinical data are multiple observations and may involve several different observations that are made concurrently. There is a broad range of data types ranging from the narrative; textual data; numerical measurements; genetic information; recorded signals; drawings; and even photographs or other images. [103]

Within EM, clinicians are central to the process of data collection and interpretation, as they converse with patients, effectively gathering narrative descriptive data. For example, finding out details about the nature of the complaint; patient history; family/social information; examining the patient; ordering and interpreting test results; considering diagnoses; devising treatment plans, and communicating with other providers about the appropriateness of admission or discharge. [104, 105]

Furthermore, patient examination and assessment usually require additional data, e.g. ordering laboratory or radiologic studies and observing responses to therapeutic interventions. Other healthcare workers may also contribute to the data-collection process:

- Administrative staff may record demographic and financial information.
- Therapists record treatment results and suggestions for further management.
- Laboratory personnel perform tests on biological samples, e.g. blood or urine.
- Radiology technicians perform X-Ray examinations, and Radiologists interpret the results.
- Pharmacists interview patients to ascertain current medications or known drug allergies.

Hence, all of these findings are recorded, and information is made available for later use by clinicians and nurses. [103] For emergency cases, additional sources of information summarising the nature of the incident, suspected injuries, on-site treatment provided by ambulance personnel or the police, will usually be provided to clinicians when handing the patient over to ED.

Information is generated when procedures are performed, and simply, by the presence and flow of patients, [104] hence, large amounts of information inevitably amass, and even more so for complex cases. Medical records allow the storage, organisation and compilation of detailed information about individual patients. [103] They support accurate and timely information collation and act as a channel to aid communication. Thus, ensuring that messages between clinicians remain well-formed

and maximising the likelihood that critical information is 'sent' and 'received'. [19] In this respect, medical information for any patient must be relatively complete to be of value for clinical use. [105]

EM is a complex and often-chaotic environment, presenting numerous challenges for personnel including paramedics, nurses, diagnostic radiographers, reception staff, porters, healthcare assistants and ED clinicians. [60] Emergency care providers often work under pressure in conditions involving disrupted sleep cycles, contend with multiple interruptions, acute time constraints, and patient overcrowding. [54] Additionally, in some instances, ED clinicians must implement medical interventions and make crucial decisions utilising limited historical or diagnostic information or even incomplete records. [54, 104]

Nonetheless, clinical decision-making is information dependent, and information and communication are deeply intertwined. [18] Communication can be defined as "The imparting or exchanging of information by speaking, writing, or using some other medium". [106] Clinical information may be obtained in a formal manner, i.e. from permanent sources such as patient notes, or it can be informal, i.e. obtained verbally from colleagues, or annotated in personal notebooks and as such may not be transcribed to a patients' medical record in great detail. [18]

Interpersonal communication is a means of information exchange [20], and communication amongst healthcare workers accounts for the majority of information flow within healthcare environments. Clinical 'messages' store information in a variety of media and formats, and can be delivered in a variety of ways that include face-to-face conversations, letters, e-mail, voicemail and medical records in either electronic or paper format. Consequently, the quality of clinical care may be affected significantly by poor information exchange, e.g. the poor presentation of clinical data can lead to poorly informed clinical practice, inappropriate repeat investigations, unnecessary referrals and waste clinical time and resources. [107]

Every clinical action, choice of treatment and investigation completed is formed predominantly on the information available, which can be thought of as the clinical evidence that informs judgement about the right course of action, as clinicians gather evidence through communication. [24]

However, failure to communicate effectively may lead to errors inadvertently resulting in adverse clinical events or outcomes, e.g. clinical morbidity or mortality. [19, 20] An error can be defined as "the failure of a planned action to be completed as intended or the use of an incorrect plan to

achieve an aim", whereas, in this instance, safety can be defined as "freedom from accidental injury". [108]

Fordyce et al. (2003), [54] undertook an observational study which attempted to categorise emergency care errors. The study reported the top five reasons, in respective order, as being due to diagnostic studies; administrative procedures; pharmacotherapy; documentation and communication, where communication had been described as a collaboration with physicians and other professionals to evaluate and treat patients.

Nevertheless, ED clinicians are subject to high levels of interruption, partly due to the interrupt-driven nature of the hospital environment. [19] A previous observational study undertaken by Coiera et al. (2002), [109] looking at communication loads on clinical staff in ED, found that the combination of interruptions and multiple concurrent tasks for clinical staff inevitably resulted in clinical errors by disrupting memory processes. Consequently, interruptions impose cognitive loads on clinical staff and have a negative impact on memory, leading to a clinical error. [20]

Additionally, emergency care providers also experience unnecessarily high communication loads, and in retrospect, many tasks may be accomplished by the simple access of formal information sources, e.g. medical records, as opposed to asking informal questions. [109] An open and non-punitive environment where information is shared freely, and responsibility broadly accepted will facilitate patient safety and worker satisfaction. [110]

Thus, within hospital settings, emergency care providers are; first, eager consumers of available historical clinical data, and second, are creators of information that is used for patient treatment and follow-up. [104] It is therefore crucial for emergency care personnel to access information, such as the patient medical record, to facilitate effective communication and inform the clinical decision-making process.

2.3.2 Information fragmentation within emergency medicine

The NHS is the largest semi-integrated health system in the world but remains fragmented in how it organises the delivery of care, and as a consequence, boundaries between the provision of primary and secondary care will be inevitable. [2] One of the greatest challenges facing EM due to its very

nature is that of fragmentation. A fragment can be defined as “An isolated or incomplete part of something”. [111]

Within the field of EM, focusing on hospital environments, A&E services frequently work in conjunction with many key departments, including radiology; pathology; observation areas (clinical decision units); pharmacy; other clinical specialties; admission units and psychiatric assessment services. [70]

Additionally, any department, system or process, which operates in isolation from others, can be said to be working in silo. [3] Organisational silos may inadvertently develop amongst numerous services via acquired specialty expertise, time pressures, and structural configuration. Although EM services are delivered via distinct service providers and departments, each department or discipline may have its own lexicon, inevitably resulting in the creation of boundaries that can inhibit information and knowledge sharing. [4]

As EM fundamentally relies upon a complex interlinking of several services, the potential for misalignment of effective communication, and failure of coordination amongst these services will be significantly increased, which can be detrimental across the continuum of emergency care. [55]

For example, an analogy between EDs and a large jigsaw puzzle can be drawn, whereby each piece (or fragment) of the puzzle is a specialist service and only when the pieces of the puzzle are put together can the entire picture be visualised, essentially showing the continuity of care. Additionally, within each jigsaw piece (specialist service) is transcribed a fragment of vital information which tells a specific part of the overall story. Assuming the same lexicon has been used, it is only after this information has been combined by the arrangement of the pieces in a logical temporal sequence will it truly amalgamate and form into a coherent story.

Nevertheless, as EM is fragmented, it follows that all information and associated storage mediums/devices contained within each service will also be fragmented. Thus, the patient medical record may well be split or subdivided amongst various departments.

In conclusion, the EM environment necessitates a higher level of both formal and informal communication. However, clinical personnel are subject to high levels of interruption, which in turn increases cognitive loads and results in the potential for errors, inadvertently affecting patient

safety. Additionally, the quality of information management may indirectly affect how well providers manage the care of their patients. [104] Nonetheless, the issue of fragmentation underlies the healthcare crisis, [30] and investigation into addressing the problem of information fragmentation is of crucial importance.

2.3.3 Medical records

A medical or patient record is a repository of information about a single patient, created by healthcare professionals following direct consultation with the patient. [112] Historically, the medical record arose in the nineteenth century as a personalised laboratory notebook, on which clinicians would record observations and plans as an aide-memoire of pertinent details for when they next saw the same patient. However, as the field of medicine has significantly evolved over the decades, so too have the requirements for healthcare provision. Correspondingly, the medical record has also been developed and adapted to accommodate these changes. [99]

Medical records store information about a specific individual and include patient demographic details; previous appointments; illnesses; comorbidities; allergies; diagnostic test results; treatment plans; medications and clinical observations. Healthcare professionals use medical records to manage patients safely, undertake audits, improve services, and potentially undertake research. Hence, the medical record is able to support effective communication between different healthcare professionals. [98]

Traditionally, ‘paper’ was used as the medium on which to store medical records, as it was readily available and portable enough to be carried to the point of care. Paper records facilitate easy annotation, allow users a high degree of flexibility in the recording of subjective data and are readily adaptable to many tasks and users. They can also be browsed through relatively easily and can be organised in various ways as appropriate. [112, 113]

However, the format of the paper record does not lend itself easily to dealing with patients that have multiple problems over a long period of time; hence, paper records may become difficult to manage and organise. Depending on the severity of the patient illness, the record may become very large and unmanageable spanning several pieces of paper, making it difficult for healthcare professionals to glean, extract or analyse the information recorded within. [112, 113]

Furthermore, for institutions such as hospitals that treat/manage a large number of patients, paper records will inevitably cause storage space issues. As paper is fragile, susceptible to damage and degradation over time, once a paper record is lost, it is forever irretrievable. Paper records can only be used for one task at a time, and access is restricted by location and office hours. Sharing paper records between healthcare professionals involves document duplication, using the fax or post, which is inefficient, wastes time, risks potential loss and has legal implications. [112, 113]

Consequently, paper-based medical records are increasingly recognised as inadequate for meeting the needs of modern medicine [98] and advancements in information technology have bought about the rise of the Electronic Medical Record (EMR), which put simply, is the computer replacement for existing paper record systems. [113]

The EMR is a computer-based record, residing within a specifically designed clinical information system (CIS) which supports users by providing access to complete and accurate data, providing alerts and reminders, with some CIS even providing clinical decision support and linkage to medical knowledge. [112, 113]

The Institute of Medicine [112] defines the computer-based patient record as “an electronic patient record that resides in a system specifically designed to support users through the availability of complete and accurate data; practitioner reminders and alerts; clinical decision support systems; links to bodies of medical knowledge and other aids”.

Computer-based systems possess several powerful physical attributes that make them ideal data capture and storage systems, as they can store enormous quantities of electronic data within a small physical space. Digital data can be duplicated easily and cheaply, e.g. back-ups for security purposes. Speech can be converted to electronic data by using dictation software, and voices can be digitally recorded and images annotated using, for example, a computer stylus type device. [113] Numerous medical data, clinical cases and advice given by clinicians to patients can be stored as different file types, e.g. medical imaging files, text documents, videos, images or even sound files. [13]

Additionally, as patient data are stored in an electronic (digital) format within database technologies, this storage medium has several advantageous features that facilitate rapid and simultaneous access to the patient record by multiple users. Network technologies further ensure that these records are

securely accessible/retrievable from a variety of locations, e.g. in instances where individuals may be geographically separated but can process information on-line.

Furthermore, computer systems can be used for automating clinical/administrative processes and expedite accurate data/information capture via using pre-defined lists of values at the data entry stage. In addition, when dealing with a large number of electronic patient records, the speed of searching through electronic data is significantly faster than manually searching through paper records. Another significant advantage of computer systems is their potential to facilitate the integration of data from multiple and disparate data sources. [113, 114]

Despite EMR overcoming many of the problems attributed to paper-based records, there remain challenges to implementing EMR at the organisation and individual level. [115] Some limitations of electronic records, as described in the literature include record content; record format; secure access; record availability and retrieval; record linkage and integration. [112]

Nevertheless, although individual computers are relatively inexpensive, the costs of maintaining large networks, employing specialist personnel and training healthcare workers may be steep. A key challenge to EMR use has been their adoption and integration into clinical workflow, and system benefits must reflect the time and effort put in by its users. [115]

The primary purpose of an EMR is to store information about the patient, which is usually generated/transcribed by physicians, nurses and hospital administrators. [8] In essence, the EMR provides a single point of access for relevant active data about a patient; acts as an informal workspace for clinical recording; maintains a medico-legal archival record of what happened to the patient; and offers an efficient means of communication between healthcare professionals. [116]

The EMR can be further subdivided into two categories, firstly the ‘Electronic Patient Record’ (EPR), which captures the periodic episodes of care provided to the patient mainly by one institution, e.g. specific treatment for a condition at an acute hospital. [10]

Secondly, the ‘Electronic Health Record’ which is a longitudinal record of a patient’s health, and can be thought of as a lifelong record ‘from the cradle to the grave’ as it provides an overview of a person’s health episodes, assembled from summaries of individual electronic patient records and

other relevant information. Within the UK, almost all of a patient's hospital medical records are copied to their GP. [10, 117]

However, no single medical record structure will suit every purpose, [118] and the environment in which the medical record is used will strongly influence the type of information recorded; the medical record layout/structure; information storage limitations and sharing policies, hence, a few variations of the medical record exist. Data within an EMR can be structured and viewed chronologically, by source, by problem or even a combination of views. GPs have a patient-oriented view of the information stored in medical records for treating patients, whilst hospital clinicians have a specific episode/case-oriented view to the medical record. [8] Coirea (2015) [118] identifies four common medical record structures:

Integrated Records are typically employed in hospital environments, within the EPR, data are presented in a strictly chronological manner, and each episode of care is identified by date and time. For example, data from investigations such as radiology may be followed by progress notes written by a clinician, subsequently followed by a change in medication orders. The integrated record is data-oriented, i.e. it provides little to no structure beyond a timestamp [118], e.g. Episode one, Episode two, Episode three.

Source Oriented Medical Records are organised according to the source that generated the data, typically different hospital departments. There are separate sections for medical notes, nursing notes, laboratory data, radiological results and so forth. Each source section may then be further subdivided according to the different types of tests, and then arranged chronologically. This type of record focuses on presenting data in a simple indexed form, similar to a telephone directory in which each section is self-contained. [118]

Protocol Driven Medical Records may be used by for patients receiving standard treatment for a well-understood condition, e.g. asthma or diabetes. A standard pre-structured template/form dictates the specific data to be obtained and recorded by clinicians and outlines the treatment plan for the patient. This type of record design helps guide users (healthcare professionals) through all necessary steps in the construction of the medical record. [118]

Problem-Oriented Medical Records have four components: a problem list; an initial plan; a database containing all patient data; and progress notes. Data are organised per patient, separately listing all problems, each of which encompasses a set of symptoms and diagnoses. The problem list is dynamic

and acts as an index to the whole record, enabling the allocation of a name, number and date to each problem. All subsequent notes, diagnostic tests, medications will be numbered according to the problem to which they relate [118], e.g. Problem one, Problem two, Problem three.

Therefore, it is important to select an appropriate format for the EMR as this will, firstly, determine how to structure, represent and capture each item of information within a system. Secondly, the information obtained can be manipulated and viewed using different orders/arrangements or grouped in different ways. To better facilitate the latter, data that has been coded and structured or broken down into simple elements readily lends itself to automatic processing. [98]

Moreover, there is significant evidence supporting the premise that when information technology is applied and deployed appropriately, the electronic health record systems put in place can enable safer and higher quality care than paper-based systems. [2]

2.3.4 Clinical information systems

The power of computers, usually measured as a function of speed and storage capacity, has increased by a factor of tens of thousands over the past few decades. [7] Thus, the rise in the use of technological innovations and computing within the field of medicine has led to the creation of Clinical Information Systems (CIS).

A CIS is a subsystem of the healthcare setting in which it operates, and is designed to store, process and exchange information, all of which are elements pertaining to medical documentation. [9] The primary function of a CIS is to facilitate patient care [10] by housing EMRs, i.e. all data and documents generated or received during the care of a patient at a healthcare institution. [9] Additionally, CIS can provide healthcare institutions with immediate information about the quality of patient care; performance metrics; finance; resource use, and facilitate the provision of medical knowledge. [9]

Within hospital settings, there are many types of CIS, and each has a specific purpose, for example, the monitoring of patient admissions or transfers; scheduling appointments; patient registration; recording laboratory results; pharmacy and drug prescription services; clinical decision support systems; data warehousing and analytics (clinical and non-clinical) and non-clinical strategic decision support systems. [119]

However, individual CIS are usually bespoke, possess unique designs, and are created to function within their specific environment. Most CIS will include nuances, which are reflective of the software company or vendor that designed/created them. As such, differences between CIS will become evident depending upon healthcare settings, local policies, workflow, type of patient seen, nature of the care to be delivered and technology both available and to be implemented. All these variations result in heterogeneity of CIS design, implementation and use, thus, making it very difficult to generalise the impact or compare outcomes between CIS. [113]

An information system is distinguished from other systems by its components, which include data and models, [21] for example, medical data can take many formats including text, images, digital voice recordings and physiological signals taken from monitoring devices. Although CISs are able to store a variety of complex patient data, this inevitably necessitates a requirement for increased physical system size, normally reflected by the increasing amounts of storage space CIS require. [101]

CISs store EMRs in an organised and structured manner; each CIS comprises specific software, hardware and possess various features, functions and information structures in order to operate smoothly within its designated environment. Thus, a CIS can be thought of as a database used within a clinical setting. A database is a structured repository for data, consisting of a collection of data tables and their associated relational database model, and usually stored within a computer system. [113] The main objective of a database is to try to present a faithful model of a body of information that reflects ‘reality’. [120]

The relational model was first proposed by E. F. Codd in his 1970 seminal paper, ‘A relational model of data for large shared data banks’ [121] and this paper is generally accepted as a landmark in database systems. Codd, a trained mathematician, used terminology taken from mathematics, principally set theory and predicate logic. [122]

A relational database is a collection of data items organised as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganise the database tables. [121] The process specifying exactly how all the tables are linked with one another is known as the relational model.

Within the relational model, all data are logically structured within relations (tables). Each table has a specific theme and is given a unique name for identification. Additionally, each table contains specific data elements, i.e. columns of data that comply with thematic table design specifications. Each row contains one value per column, and tables are linked using a primary key.

For example, referring to Figure 2.6 (page 68) the patient table only contains data elements that record demographic details about the patient; and each patient is identified by the allocation of a unique patient ID or primary key. Correspondingly, the 'Episode Table' contains only details about the patient's episode of care, e.g. triage notes, and the Medication Table only include details about what medications the patient received. All three tables are linked using the patient ID primary key. The relational model ensures that data are not duplicated within or across each table. Additionally, it is possible for one patient to have one or more associated episodes, or zero, one or more medications.

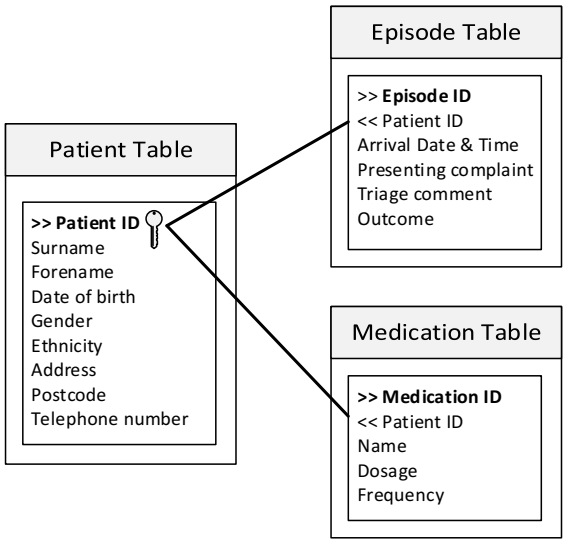


Figure 2.6 Simplified example of the relational model

The regular and formal indexing structure permits rapid retrieval of individual data elements (e.g. Patient ID, ethnicity or postcode) within each table of the database. [123] Furthermore, the relational model allows a high degree of data independence; for example, application programs should remain unaffected by modifications to internal data representation, e.g. the re-ordering of records. Likewise, to provide substantial grounds for dealing with data semantics, consistency, and redundancy problems Codd's article [121] introduced the concept of normalised tables, which in essence, sought to eliminate repetition/duplication of data within tables. [122]

Referring to Figure 2.7 (page 69) healthcare workers (clinical and administrative) will access a computer terminal connected to the hospital network, on which the specific departmental CIS is located. As such, users access the 'front-end' of the system, which facilitates live data entry and viewing of individual patient records. From a technical perspective, all transcribed data is stored within multiple tables, e.g. patient demographics table, hospital episode table or medication table, all of which are linked using the unique system primary key (patient ID).

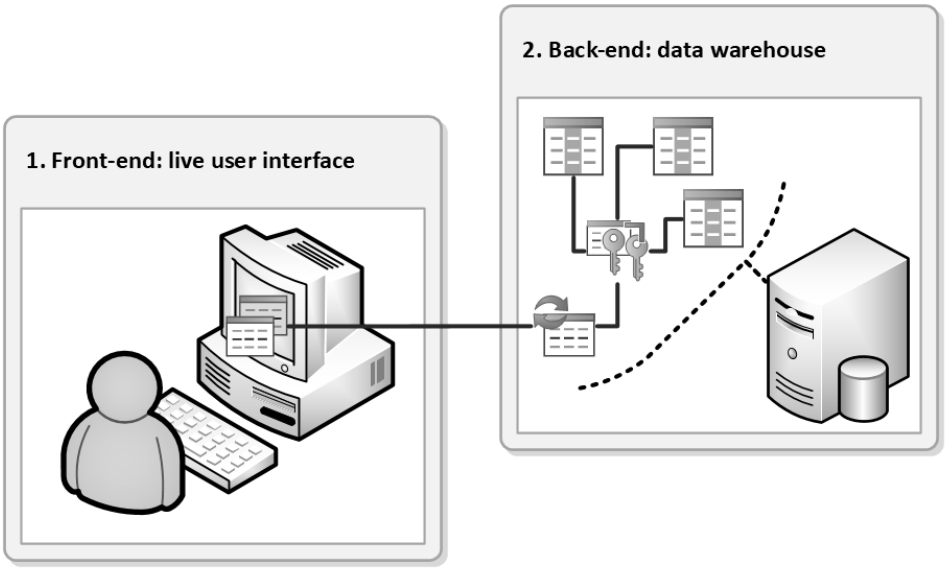


Figure 2.7 Data entry and storage within clinical information systems

Conversely, database managers and information analysts seldom access the front-end interface, typically accessing the database from the back-end using specialist software, i.e. 'Relational Database Management System' (RDBMS) to view the collection of relational tables from the back of the CIS. RDBMS has become the dominant data processing software in use today. [122]

A typical database may contain hundreds of tables, and to facilitate navigation and location of specific data elements within the relational model of a database a data dictionary is used. Data dictionaries are often live documents created by database developers that detail every specification of all data elements within tables.

For example, when looking at the data element 'Patient Arrival Date/Time', the dictionary would contain a description, stating this was the Date/Time a patient first presented to a department. Secondly, the format Day/Month/Year/Hour/Minute/Second would be specified. Additionally, data dictionaries provide a roadmap allowing other technical users to understand/navigate the complex relational structures between tables.

At the heart of the CIS lie patient data, and the configuration in which CIS have been designed and organised is essential to the functioning of the system. [124] Even though clinical data are naturally hierarchical, designing/developing a structured patient database containing comprehensive clinical information is a challenging task within any database model. [124]

Relational databases store a wealth of data and information, which requires secure access and can be queried to facilitate analysis and support an organisation's operational functions; this can be achieved via RDBMS software. The management and manipulation of tables within a relational database/model have led to the development of a structured query language called SQL, which is flexible, and able to accommodate potential requirements into looking at or querying data in new ways. SQL has since become the formal International Organisation for Standardisation and de facto standard language for RDBMS. [122, 124]

SQL facilitates communication between humans and databases [101] by permitting the questioning/querying of several relational tables within one script. For example, referring to Figure 2.6 (page 68) the SQL to identify all female patients with the surname 'Brown' would be written as:

```
SELECT      PatientID, Surname, Forename, DateOfBirth, Gender, Ethnicity
FROM        PatientTable
WHERE       Gender = 'Female' AND Surname = 'Brown';
```

Nevertheless, the quality of information may inadvertently affect the level of care a patient may receive, as every clinical decision is predicated on information about the patient and about the treatment options available. Whenever a decision is important, complicated, critical or made often, the introduction of an information system will help to support/streamline the decision-making process. Furthermore, in the case of complex or critical decisions, the information system may also help to reduce complexity or minimise the likelihood of an error. [21, 22]

Within a hospital environment, CIS have the potential to help transform the delivery of care via the re-designing of disease and care pathways; enabling better coordination of care through information sharing; utilising data in improving quality; safety and efficiency of care provided; clinical decision-making and informed business intelligence. Furthermore, CIS possess the capacity to enhance coordination and integration between primary, secondary, specialist and social care. [2]

2.3.5 Disparate clinical information systems

Within NHS hospital settings, there are a multitude of departments and clinical specialties which need to exchange information. [7] However, each department houses bespoke CIS, which although perform optimally, only record information specific to the functionality of the department. As patient information is complex, there is no 'one' overall CIS designed to offer access to all the available information within a hospital. [8]

Additionally, it would be difficult for single CIS vendors to have expertise in all types of systems, especially in niche systems specifically designed for a single function, e.g. neurology. [125] Hence, within a single health system, patient information may be difficult to access due to the use of different CIS within, e.g. ambulatory, inpatient, and emergency settings. [11]

Thus, attempting to integrate multiple legacy systems with new systems would present many challenges that prevent integration. As specialised CIS are designed to address the requirements for specific departments, clinicians/healthcare staff may also need to transcribe the same information within the main Hospital CIS, for the cohesiveness of the patient record. [125]

Furthermore, CIS within specialty departments are often installed/implemented at different times over many years, so the functionality is in a state of constant flux. Hence, many organisations have disparate CIS in use, some include outdated 'legacy systems', whilst others feature newer technologies. [125]

Therefore, the majority of CIS are explicitly designed/developed to perform specific tasks within departments by different vendors, using different technologies or computer languages. [12, 13, 29] Hence, organisations using multiple disparate CIS created by different vendors inevitably face challenges, e.g. disparities in system standards; formats; system interfaces/platforms; operating systems; database management systems; network protocols; data standards; data formats (e.g. structured/unstructured); data sharing capabilities; regulatory compliance; vendor dependency; vendor lock-in; multiple system costs; increased staff training requirements, all of which obstruct the sharing of data. [12, 13, 29, 126]

Consequently, medical records concerning the same patient that were transcribed in different CIS and allocated unrelated identifiers will inevitably result in fragmentation of the patient record, which in turn will lead to future complications when trying to re-identify the patient, ultimately triggering

the formation of the data silo. Most healthcare data, whether on paper or electronic format, are trapped in 'silos' inadvertently across health systems, e.g. between primary care and secondary care, or even within the same hospital between different departments. [5]

As increasing amounts of heterogeneous medical data become available, healthcare organisations and hospitals face many challenges such as technical, security and human interaction in freeing data from silos, inadvertently affecting both patient care and medical research. [15] In addition, fragmented patient information risks patient safety; hinders data aggregation for clinical decision support; prevents physicians from having comprehensive medical information; deters effective population health approaches; creates inefficiencies by delaying care; limits public health reporting; and severely reduces the utility of electronic data for clinical research. [17]

Ideally, the various CIS within a healthcare organisation should work in a cohesive manner, communicating with one another, interchanging data and performing optimally. However, in reality, this is seldom the case, since most of the proprietary software systems created by different vendors do not communicate with each other. This highly heterogeneous and distributed nature of healthcare computing only serves to exacerbate information communication problems. [127]

Additionally, medical data and information within disparate CIS can be stored and represented in different ways, leading to discrepancies. For example, when coding the data element 'Gender', one system may code the data as Male = 1 and Female = 2, whereas in another system the code could be abbreviated to Male = M and Female = F. Another example is differing data structures, looking at 'Date' elements where the format could be stored as either 'DD-MM-YYYY' or 'YYYY-MM-DD' or 'MM-DD-YYYY'. [29]

Thus, CIS may become disparate or isolated due to a lack of interoperability with other CIS. To facilitate the sharing of data, interoperability is the only sustainable way for enabling CIS to talk to one another in order to obtain the complete 'image' of a patient. [15]

Nonetheless, any decision to purchase a CIS from a vendor offering a less-interoperable system will be counterbalanced with the need to exchange electronic information. However, it remains feasible that healthcare organisations may have procured legacy systems before interoperability concerns were as important as they are today. [128] Nonetheless, without external pressure from the marketplace, clinicians or healthcare staff, vendors have little or no incentive to change. [129]

2.3.6 Interoperability

The Institute of Electrical and Electronics Engineers (IEEE) defines interoperability as: “The ability of two or more systems or components to exchange information and to use the information that has been exchanged”. [16] There are two classes of interoperability [130]:

- **Syntactic** (technical or functional) interoperability is the process of moving data between two systems (i.e. ‘how’). It can be thought of as the ability of computers systems to exchange data, without necessarily understanding what is being exchanged. For example the USB specification for attaching peripheral components such as keyboard, mouse or printer. [7, 130]

- **Semantic** interoperability is the process of ensuring each system can understand the information received from others (i.e. ‘what’). Hence, it is the capability to share data between applications and is dependent upon both sender/receiver understanding the common language. For example, English is used a common language for air traffic control, which does not mean that all systems need to use this language internally, only that it is used to facilitate understanding between different countries, i.e. to facilitate interoperability. [7, 130]

Additionally, the Commission of the European Communities [131] define interoperability from an EMR perspective:

> *"Interoperability of electronic health record systems" means the ability of two or more electronic health record systems to exchange both computer interpretable data and human interpretable information and knowledge.*
>
> *"Semantic interoperability" means ensuring that the precise meaning of exchanged information is understandable by any other system or application not initially developed for this purpose.”*

The authors Benson et al. (2016), [132] describe interoperability as having four layers, each of which is further subdivided, as summarised in Table 2.7 (page 74). Each layer must consider the issue of interoperability firstly, within its own context, and secondly, in relation to other layers.

When considering semantic interoperability, there remains the potential of electronic messages sent between two CIS to be misunderstood, i.e. the message that was sent, may not be the message that was received, due to unintentional modification by the chosen communication channel. Thus, the effectiveness of communication between two CIS is dependent upon the communication channel

employed, which will inadvertently vary in its capacity to carry data. Hence, data message transference is susceptible to transmission delays, noise or distortion. [24]

Table 2.7 Layers of interoperability (adapted from Benson et al., 2016)

Interoperability layer	Subdivisions
Technology	• Syntactic or technical interoperability • Exchanging data between two CIS • Interchange formats • Application Program Interfaces (API) • Health Level 7 *
Data	• Semantic interoperability * • Exchange meaning between two CIS • Identifiers and Codes • SNOMED CT
Human	• Process interoperability • Clinical interoperability • Business processes interoperation
Institutional	• Role of Government • Culture • Incentives • Education • Regulation • Information governance

** Detailed within Section 2.4 (page 79)*

Furthermore, from a data perspective, the authors Liyanage et al. (2015), [23] state the heterogeneity of data may be observed at three levels, firstly, semantic which refers to different ways of interpreting the meaning of data. Secondly, syntactic which refers to different ways of formatting the data, and thirdly, structural, i.e. referring to the different ways of storing data.

Another critical factor is the heterogeneity of the CIS or platforms that host data [23], as these determine the 'knowledge' possessed by each CIS and any common ground between them. [24] Thus, the most commonly encountered difficulties include CIS resource limitations; cognitive limits on memory; the system's ability to complete the task at hand in a timely manner [24]; data heterogeneity; data complexity; noise and incompleteness [15, 23], all of which act as barriers to interoperability.

Likewise, Gagon (2007), [133] reports that heterogeneity conflicts arise due to first, Syntactic heterogeneity, which relates to differences between data models, i.e. entity-relationships, relational,

object-oriented. Second, Structural (or schematic) heterogeneity in which different information systems store their data in different structures, and third, Semantic (or data) heterogeneity which considers the content of an information item and its intended meaning.

Ultimately, the degree of interoperability between two systems is affected by the limitations of the communication process and of the CIS sending and interpreting the messages. [24] Hence, semantic heterogeneity is often cited as the biggest challenge that needs to be addressed within the context of specific domains. [23]

Moreover, as CIS are proprietary, such issues can only be addressed with collaboration between vendors, and in relatively few cases CIS vendors have collaborated to enable electronic data exchange for a limited range of tasks such as reporting clinical test results. [7]

Interoperability is a key issue and a long-term domain of research for distributed healthcare information systems. [29] The lack of interoperability has been an impediment to progress for many years [7], hindering the process of creating value or gaining insight from data. [15] Benson et al. (2016), [132] identify the most common difficulties to interoperability encountered by various perspectives, summarised in Table 2.8 (page 75).

Table 2.8 Common difficulties relating to interoperability (adapted from Benson et al., 2016)

Perspective	Details
Bit-perfect	• Computers work by matching strings • Need for perfection • Specifications must be stringent
Translation issues	• Computers store data in different ways • Double translation (using intermediary wire, e.g. JSON or XML) i.e. FROM CIS A to wire format, THEN wire format to CIS B • Need a standard interchange language all systems can use
Developers	• Each system speaks its own language • Managers wish to re-use existing software • Lack of domain knowledge • Developers and users fail to understand each other
Users	• Each specialty speaks its own dialect • Users do not know what they really want • Users won't commit to requirements • Users keep asking for changes • Users do not understand software development

A report from the National Advisory Group on Health Information Technology in England highlighted that “the digitisation of hospitals has been far from smooth, and the patchy computerisation of this sector stands as a considerable impediment to transforming care”. [134]

A recognised example of the complexities in addressing health system interoperability is the National Programme for Information Technology (NPfIT). In the mid-2000s, NPfIT was an ambitious £12.4 billion Government investment designed to reform how the NHS in England used information to improve service and patient care. The aim was to enable the joining up of information systems and datasets and create a single centrally mandated electronic care record for patients, which would facilitate the rapid transmission of information between different parts of the NHS. This was hoped to improve patient outcomes and better inform the direct provision of care throughout the NHS. [2, 134]

However, NPfIT did not achieve its aim because of multiple factors which predominantly included failing to digitise the secondary care sector. In addition, as NPfIT was too centralised, it failed to engage properly with Trusts and their respective healthcare professionals, whilst trying to accomplish too much too quickly. [2, 134]

Nevertheless, enabling seamless health information exchange, i.e. a computer network designed to allow different CIS to interconnect and share patient-level data remains the target. [123] There remains a major need for the integration of overlapping information being held within heterogeneous CIS, even if they have been developed at different times, by different vendors using different technologies. [8] Thus, an efficient solution to incompatibility of data exchange is of crucial importance. [15]

2.3.7 Medical informatics

Healthcare is an information-based science, and much of clinical practice involves the gathering, synthesising and acting upon information. [26] Nevertheless, it is plausible that health information is not being utilised effectively; hence, a range of solutions will be required to address such information-related problems within healthcare settings. [115]

Medical Informatics (also known as Health Informatics) is the field concerned with the management and use of information in health and biomedicine. [115] It encompasses the study and application of

methods to improve the management of patient data, clinical knowledge, population data, and other information relevant to patient care and community health.

Warner (1995) [135] defines Medical Informatics as 'the study, invention, and implementation of structures and algorithms to improve communication, understanding, and management of medical information'. The author further identifies five categories which seek to outline what Medical Informatics encompasses, i.e. Signal processing; Database design; Decision-making; Modelling and simulation, and Optimising the interface between the human and the machine.

Moreover, Medical Informatics can be thought of as a multidisciplinary field, encompassing elements from the intersection of clinical sciences, the public health sciences (e.g. epidemiology and health services research) as well as cognitive, computing, and information sciences. [26] Medical Informatics is the study of information/communication processes and systems within a healthcare setting, and focuses on [123]:

1. Understanding the fundamental nature of information and communication processes and describing the principles that shape them.
2. The development of interventions that can be used to improve existing information and communication processes.
3. The development of methods and principles that allow such interventions to be designed.
4. Evaluating the impact of such interventions, on the way organisations and individuals work and on the outcome of their work

Thus, Medical Informatics seeks to support clinicians with their decisions and actions and to improve patient outcomes by making better use of information [98] by improving the efficiency in the way in which patient data and medical knowledge are captured, processed, communicated, and applied. [98]

Many of the techniques and solutions involve the application of computers and computer-related technologies [115] which ultimately enable healthcare professionals to use information to improve patient health outcomes. [14]

A number of core themes underlie Medical Informatics, including standards, terminology, usability, and demonstrated value. There is a high degree of interaction amongst a federation of different

computer systems from different vendors; therefore, standards are essential to facilitate the integration of data across systems. As such, standardising the structure of clinical documents will serve to simplify aggregation and movement of data across systems. [115]

Nevertheless, for any Informatics based approach numerous factors will need to be taken into consideration such as the specific problem that is to be addressed; s the application of technology the best solution; or are there simpler alternatives?; whilst bearing in mind the possibility, practicability and desirability of the chosen approach.

Therefore, it is important to consider the following questions theoretically: [123]

- What can be actually be achieved given the constraints of the real world?
- What is the potential for successfully engineering a system or introducing a new process?
- What is the fundamental motivation for using a given process or technology?

2.3.8 Section summary

This section provided an overview of what is health information and how it can become fragmented within an EM environment. It described the patient medical record, focusing on the types of Electronic Medical Records, i.e. 'Electronic Patient Record' and 'Electronic Health Record'. An overview of Clinical Information Systems was provided, within the context of a hospital setting and the inadvertent prevalence of multiple disparate CIS resulting due to a lack of syntactic and/or semantic system interoperability. Additionally, a succinct overview of the field of Medical Informatics was given in order to provide appropriate context for this research.

2.4 Data integration

Data integration can be thought of as the process of combining disparate data and providing a unified view of these data, by relating otherwise separate data and information resources. [14, 123] Both academic and practitioner communities recognise the challenges of integrating healthcare to improve service delivery to patients. [119] The vision of integrating information from a variety of sources into the way people work, to improve decisions and process, is one of the key driving factors for technological advancements in the field of informatics. [14]

Within heterogeneous CIS, access to medical information is dispersed amongst multiple proprietary information systems. [12] Each CIS is managed individually, retaining its own relational data model, database, business logic, and essentially functions like a discrete or independent unit. [14] A collection of databases does not necessarily imply that they will integrate readily.

Furthermore, CIS integration interface varies significantly between systems, especially legacy systems. [12] Thus, system interfaces may not always be compatible, and in some cases, CIS vendors may not wish to divulge details nor collaborate with one another due to rival competition, intellectual property, time invested, cost and adapting their own systems to accommodate connection with another, consequently leaving the medical information locked in silo within a CIS. [12]

System interfacing approaches to connect disparate systems conventionally include inter-systems communication techniques or a common enterprise information architecture, which can be used to connect islands of enterprise systems through the use of enterprise application integration methods. [116] Additionally, customised integration processes increase overall cost due to the additional integration time and maintenance required for modification. [12]

Integrating heterogeneous CIS presents its own challenges for healthcare organisations. [12] There are two key dimensions of systems integration, firstly internal integration which is the degree of integration between systems and technologies within an organisation. Secondly, external integration, which is the degree to which systems and technologies interface with computer systems outside the organisation. [136]

Moreover, a system may be considered to be a set of interrelated elements, either individual CIS or sub-systems, and configuration of these systems depends upon the nature of these interrelationships. [137] The field of Information Systems identifies four domains of Clinical Information Systems integration, i.e. Technical, Systems, Strategic and Organisational. [138]

First, the **Technical domain** seeks to make sophisticated software and hardware artefacts communicate via appropriate protocols, conventions and technologies. For example, the sharing of rudimentary signal and data message content, as well as the communication of shared semantic meaning through representation in databases and application systems. [138]

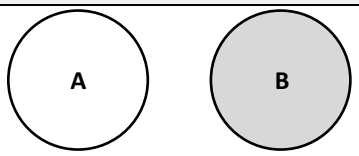
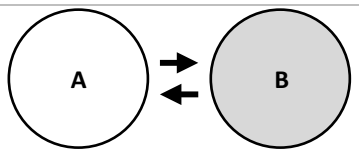
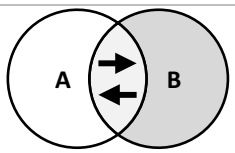
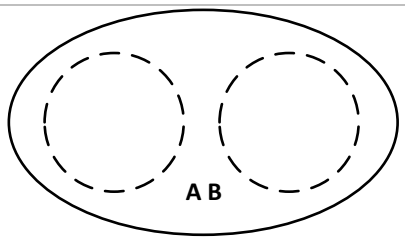
Second, the **Systems domain**, which focuses on understanding the behaviour of the technical, strategic and organisational system in providing an all-inclusive approach. As each system is complex and has distinctive adaptive and emergent properties, tasks and processes cannot be easily understood, nor dealt with in individual/discrete bundles. Therefore, a whole systems approach to the interactions between individual systems must be considered. [138]

Referring to Table 2.9 (page 81), which outlines four types of interrelationships, as identified in an article by Das (1992). [137]

- Within a **stand-alone system**, elements make their own decisions and do not communicate with any other element.
- An **interfaced system** is one whose elements have one-way or two-way communication with other elements, but make decisions for personal benefit. Thus these systems can be thought of as having a parasitic relationship.
- An **integrated system** is one whose elements have two-way communication with other elements, and make decisions for the collective benefit of the system, both systems having a symbiotic relationship.
- The **universal system** encompasses a group of elements with no individual decision-making capability and a central controller.

Thus, successful systems integration will result in greater coherence, co-ordination, management control and overall productivity. [138]

Table 2.9 Different types of system interrelationships, adapted from Das, 1992.

Interrelationship	Visual representation	Description
Stand alone		▪ Elements make their own decisions ▪ No communication between elements
Interfaced		▪ Make decisions for own benefit ▪ One or two-way communication
Integrated		▪ Make decisions for combined benefit ▪ Two-way communication between elements
Universal		▪ No individual decision-making ▪ Centralised control ▪ Single database

Third, the **Strategic domain** considers the goal of system integration to be simplified by ensuing careful strategic planning and financial investment to enable large-scale systems integration within organisations, i.e. between externally focused and internally functional areas. [138]

Fourth, the **Organisational domain** encourages the development of a better understanding of the many issues involved in system integration and advocates the introduction of appropriate policies and procedures to mitigate this task. Each system implementation is unique to its context; for example, the effect of introducing a new department computer system may resonate with other services within the organisation and affect workflow. Additionally, staff may be expected to work with new departments and share or process information in new ways. [138]

To date, the technical and system approaches utilised for integration have been mostly adapted from generic system integration methodologies. [119] Integration is not a straightforward process, and various dimensions must be considered, which take into account overall workflow, data structures, recording, extraction, exploration and homogeneity. [14]

Integration aims to bring parts together or combine them into a whole, whereas relation establishes an association, link or connection. As such, the objective should be to seek an equilibrium between data integration and establishing links or relations between data. [14]

Vast quantities of data are generated through the healthcare process, and technological advancements such as EMR and CIS facilitate the capture and accumulation of information about patients and their medical conditions, providing clinicians with the point-of-care tools they need to deliver better patient care. [14, 114, 139]

Moreover, the value of patient-specific information lies in the aggregation of individual data. Research and evaluation of stored clinical data may lead to the discovery of trends, patterns or relationships hidden within the data, which have the potential to provide new medical knowledge and enhance understanding of disease progression and management. However, there have been few tools or methodologies developed in order to discover, evaluate and analyse this hidden knowledge. [14, 139]

Nonetheless, extracting patient-centric data held within the EMR from one or multiple CIS in a secondary care setting is often an ad-hoc process that is poorly described in the literature. [140] Therefore, a solution to the incompatibility of health data exchange is of crucial importance. [15]

2.4.1 Ontologies

The classic approach to retrieving information resources relies upon the use of a controlled vocabulary consisting of ‘keywords’ or terms by which the content has been pre-indexed. [141] For example, the National Library of Medicine’s MEDLINE (MEDLARS Online) database, available via the World Wide Web, contains a controlled vocabulary ‘Medical Subject Headings’ (MeSH), [142] which consists of terms used to index each reference to the medical literature, facilitating medical literature searches. [143] For example, looking at the MeSH Descriptor Data for the term ‘Alzheimer Disease’ (Figure 2.8):

The MeSH terms are arranged hierarchically into trees covering a number of topic axes, and the tree structures are ‘poly-hierarchic’, i.e. a main heading can appear in several MeSH subtrees. The MeSH tree is an example of a ‘taxonomic’ system in which controlled ‘terms’ are used, but these terms do not represent unique concepts i.e. the same term can occur in multiple subtrees of the MeSH hierarchy. [141, 143]

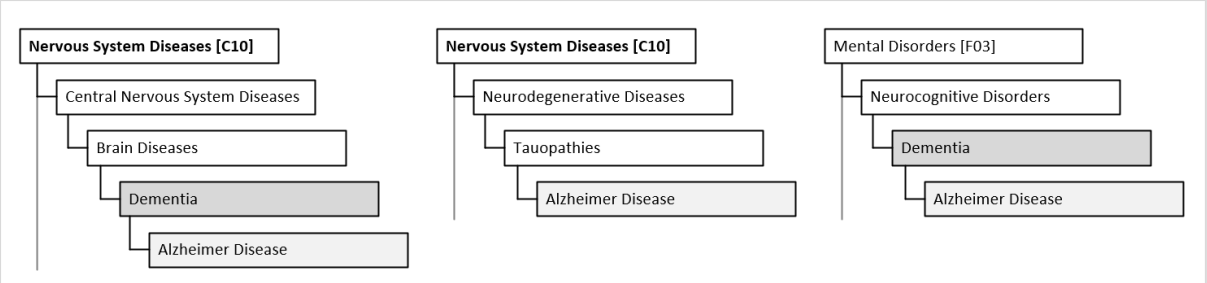


Figure 2.8 Hierarchical MeSH tree structure for term Alzheimer Disease

A hierarchy (also known as Directed Acyclic Graph) refers to the parent-child relationship of nodes that make up an object, e.g. a family tree, which provides a breakdown by classification or taxonomy. Hence, entities are organised from one or more roots 'downwards' to their children and descendants; however, no entity can be a descendant of itself. [144] Thus, 'taxonomy' refers to a hierarchical classification of controlled terms to provide a conceptual framework for a domain of interest, e.g. to facilitate information retrieval, organisation and analysis. The hierarchical relationships may be considered as 'parent-child' or 'is-member/part-of'. [141]

An ontology is analogous to a taxonomy, in that, it uses a 'concept identifier' or 'controlled vocabulary' to formally represent concepts that describe objects and associations amongst them. However, when compared to taxonomies, ontologies use richer semantic relationships to describe concepts and their attributes and possess a set of formal rules and constraints about how relations are specified, and terms are defined. Therefore, ontologies have the potential to offer a representation of knowledge, as opposed to being a method to control the terms used for the description of a domain. [141]

The word "ontology" originated within Aristotelian philosophy as the 'study of being' or simply 'of what there is' [144] Ontology, as defined within the Oxford Dictionary of Computer Science [145] is:

> *A description of some concepts and their relationships, for the purpose of defining the ideas sufficiently to allow a computer to represent them and reason about them. Thus an* ****agent****'s ontology specifies the basic building blocks of knowledge that defines what it can perceive and reason about. This is a kind of model and, as such, is very useful to define what agents or learning programs can know and what they can communicate. Ontologies are usually compiled for a particular 'domain', e.g. the domains of Medical diagnosis or Wind engineering, but they are more formal than domain knowledge.*
>
> **** Agent****: An autonomous system that receives information from its environment, processes it, and performs actions on that environment.*

Ontologies are a conceptual framework of the structured representation of a given domain or meaning [146], and they provide a method for describing concepts and relationships within a domain. [147] A conceptualisation is an abstract or simplified view of the world that we wish to represent for some purpose. [148]

Thus, Gruber (1995), [148] states 'An ontology is an explicit specification of a conceptualisation' and identifies some benefits of applying ontologies to permit the use of knowledge within computer applications; facilitate numerous computers to share knowledge; help humans to better understand a knowledge area and to construct consensus related to a knowledge area. [148]

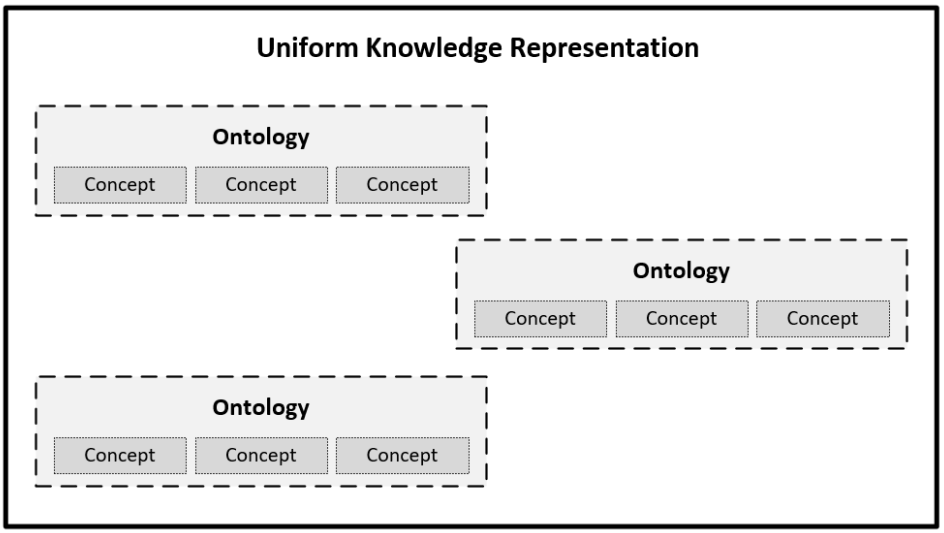


Figure 2.9 Ontologies populated by concepts (adapted from Jakus et al., 2013)

Jakus et al. (2013), [149] illustrate grouping related concepts into ontologies (Figure 2.9, page 84). The authors describe an ontology as a convenient means of connecting a subject, a relationship, and an object together. Ontologies permit the organisation of concepts under a common specification in order to facilitate knowledge sharing, and therein lies their value. For example, when considering the abstract concept of a 'Person' via ontologies, data type properties, e.g. First Name, Last Name, Gender, Date of Birth, and Home Address, can be shared with other users and interpreted in the same uniform way. [149]

The concept of ontologies has been applied to the field of Information Systems (IS) to help address issues within conceptual modelling, and the term ontology has commonly been utilised in two ways, resulting in ontologies **of** IS and ontologies **for** IS. [150]

Firstly, ontologies **of** IS, which focuses on the creation of ontologies and asks questions such as: What are the best methods to create representations of reality? How can we be sure that our ontologies are good? Although numerous ontologies have been created to describe different aspects of reality, Information scientists are primarily interested in ontologies that describe IS. As such, these ontologies will support the validation of tools used to create conceptual models. [150]

Secondly, ontologies **for** IS, which focuses on the use of ontologies, i.e. from the set of all ontologies, selecting those that have been captured in computational form and may be used to generate or validate IS components. [150]

Furthermore, Guarino (1998), [151] states "an ontology refers to an engineering artefact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words". They drive all aspects of an ontology-driven IS and are created/formalised as the first step in the system development process, after which they influence and drive the construction of the conceptual schemas and associated applications. [150, 151]

Each (symbolic) IS possesses its own ontology, whereby each symbol has been ascribed meaning according to its specific view of the world. [151] Guarino (1998),[151] further categorises these according to the level of generality of the content, illustrated in Figure 2.10 (page 85) and described in Table 2.10 (page 86)

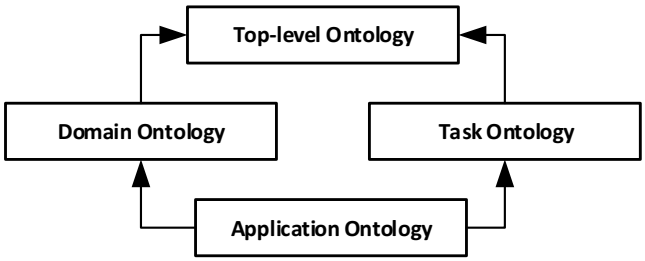


Figure 2.10 Types of ontologies (adapted from Guarino 1998)

Table 2.10 Types of ontologies (adapted from Guarino 1998)

Ontology	Description
Top-level	Describes very general concepts, e.g. space, time, object, event, action, etc. Concepts are independent of a particular problem or domain. Feasible to have unified top-level ontologies for large communities of users.
Domain	Provides the vocabulary related to a generic domain, e.g. medicine Specialising terms introduced in the top-level ontology.
Task	Describes the generic task or activity, e.g. diagnosing Specialising terms introduced in the top-level ontology.
Application	Describes concepts dependent upon a particular domain and task These are often specialisations of both related ontologies. These concepts often correspond to roles played by domain entities while performing a particular activity.

Ontologies provide the terminology/vocabulary for the information schema and rule base, whereby the rule base draws upon other forms of background knowledge. Thus, ontologies represent 'universal' knowledge that is true in all possible 'worlds' and are therefore deemed static. In contrast, contingent knowledge, e.g. databases and associated rules, pertain to just one particular 'world', hence are deemed dynamic. [144] Terminologies are comparable to a formal language in that they contain four key aspects, described in Table 2.11 (page 86).

Table 2.11 Aspects of terminologies similar to languages (adapted from Rector et al. 2014)

Aspect	Detail
Symbols or Identifiers	Concept Identifiers or the essential elements of the language
Syntax	Rules for what constitutes a valid expression consisting of: • Grammar: complete description of the syntax • Parser: program using grammar to unpack expressions
Semantics	Rules for parsing and interpreting expressions and drawing inferences Each representation language has own semantics but may have several different syntaxes, e.g. UML class models may be written in diagrammatic syntax or XML syntax.
Linguistics	How entities in the language are presented in human-readable form. E.g. in biomedical terminologies, names are labels which are invisible to computer programs, although essential to users.

Ontologies specify regimentations/arrangements of terminology that support the reusability and integration of data, thus supporting the development of systems such as decision support, information retrieval, data annotation, and natural language processing. [152]

Additionally, an ontology facilitates the development of formal vocabularies to describe basic premises of a given domain; provides a shared conceptualisation expressed in formal logic; simplifies communication between people and computational agents; promotes interoperability among organisational systems, and may even be able to replace human beings with computational agents for process/distributing type tasks. [153]

Within the field of Health Informatics, ontologies facilitate communication between humans and machines by defining the terms used to describe an area of knowledge. [147] For a defined application domain, ontologies are collections of formal machine-processable and human-interpretable representations of the domain entities, and the relations amongst those entities. [152]

Previous research undertaken by the authors de Lusignan et al. (2011), [147] identifies ontologies as being comprised of: firstly, classes (or concepts) which encompass general types of entities within in the domain. Secondly, relationships that exist amongst/ between entities within the domain. [147] Thirdly, the attributes/properties that these entities may have. [148] Figure 2.11 (page 87) adapted from Jakus et al. (2013), [149] provides an illustration to this effect.

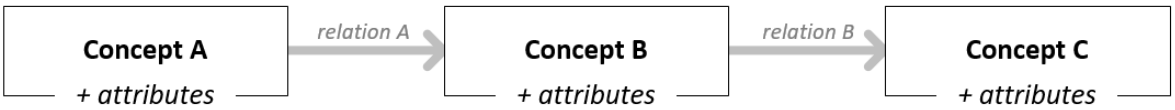


Figure 2.11 Organisation of concepts, attributes and relations (adapted from Jakus et al., 2013)

Furthermore, ontologies show great potential in achieving interoperability within a healthcare domain, as they firstly, ensure a common understating of information and secondly, make exact domain assumptions. [15] The authors Liaw et al. (2014), [154] report that ontologies can facilitate various types of integration:

Table 2.12 Types of integration facilitated by ontologies (adapted from Liaw et al., 2014)

Type of integration	Description
Data	From disparate clinical data sources within and across electronic health records.
Knowledge	From diverse health and social professionals' knowledge.
Clinical	Linking clinical concepts to model the phenotype.
Interdisciplinary	Bringing together multiple disciplines to model/support multidisciplinary coordinated care & information exchange over the patient's journey through a complex ecosystem of clinical/social factors & contexts.

Moreover, semantic interoperability is appropriate for environments in which there are known data structures, and there is transferrable meaning. [23] Thus, ontologies that enable unified information exchange across healthcare domains may be utilised to address semantic interoperability issues [155] by communicating the meaning of similar concepts used within the domain.

Table 2.13 (page 88) adapted from Liyanage et al. (2015) [23] provides an overview of the most common interoperability issues and corresponding ontological approaches suited to address these issues.

Table 2.13 Comparison of semantic interoperability and ontological approach (adapted from Liyanage et al. 2015)

Issue	Detail	Semantic interoperability	Ontological approach
Meaning	Contextual or Linguistic	Definite meaning	Infer meaning
Granularity	Completeness	Single coding system	Multiple poly-hierarchical coding systems
Temporal	Terminological changes & Definition changes over time	Consistent meaning	Evolution of meaning
Structural	Categorisation & Positioning errors	Structured records	Dynamic records

Nonetheless, even within the same information domain 'meaning' is often contextual, and when working with multiple heterogeneous data sources that are 'loosely' defined, such standards for semantic interoperability become increasingly difficult to apply. [23] The multidimensionality of the healthcare domain and its conceptualisation inevitably results in the creation of different ontologies with contradicting/overlapping parts. [15]

Consequently, semantic interoperability has a limitation in that it is not entirely able to integrate large heterogeneous datasets. [156] Correspondingly, messaging technologies have been unable to

attain semantic interoperability between data held within silos. [155] Additionally, ontologies represent large complex information domains, which can be time-consuming to create and maintain for respective domain experts. [23]

The authors Wache et al. (2001),[157] reviewed the use on ontologies within the context of integrating heterogeneous information sources and stated, to achieve semantic interoperability, the meaning of the information that is interchanged has to be understood across all systems.

Additionally, Goh (1997) [158] proposed semantic conflicts arise due to first, Naming whereby the naming schemes of information differ significantly, e.g. homonyms and synonyms. Second, Scaling due to the utilisation of different reference systems used to measure values, e.g. different currencies, and third, Confounding, when information items appear to have the same meaning, but in reality, they differ, e.g. different temporal contexts. [133, 157, 158] Table 2.14 (page 89) adapted from Gagnon (2007), [133] provides further detail on semantic conflicts.

Table 2.14 Semantic conflicts (adapted from Gagnon, 2007).

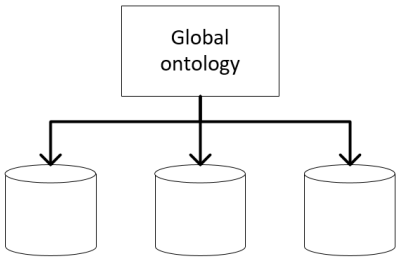
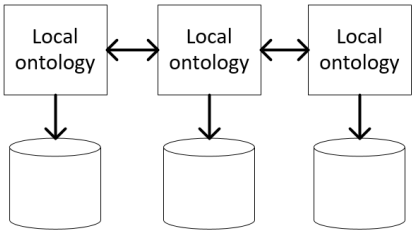
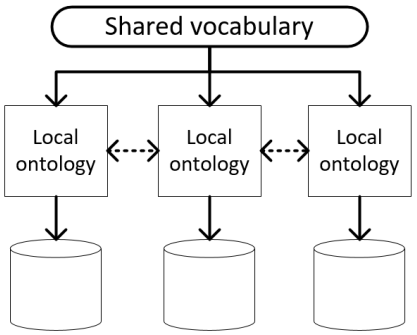
Semantic conflicts	Conflict	Description
Data-level	Value	Different meaning
	Representation	Different format
	Unit	Different measurement units
	Precision	e.g. grade and granularity differences
Schema-level	Naming	Different labels, entity/attribute homonyms & synonyms
	Entity-identifier	Different identifiers
	Isomorphism	Different attributes
	Generalisation	Different concepts
	Aggregation	One attribute for many attributes
	Discrepancies	Different structures

Furthermore, Wache et al. (2001),[157] proposed three main categories for the classification of ontology-based approaches to integration as described and illustrated in Table 2.15 (page 90).

Nonetheless, within the context of Emergency Medicine, the adoption and utilisation of ontologies have been relatively slow, for example, looking at recording the 'Patient's presenting problem' within a CIS located in an ED, which is a hectic environment where staff are subject to time constraints. Ontologies require staff to browse existing vocabularies, leading to an increased potential for errors due to incorrect selection, while also transcribing redundant patient data.

Additionally, healthcare staff may perceive that 'information' based systems cannot accurately capture the complexity of reality, so may prefer the simplicity of unstructured data entry. [159]

Table 2.15 Ontology-based integration approach (adapted from Wache et al, 2001).

Ontology-based integration approach	Visual representation
Global/Single ontology approach An integrated ontology that describes data in all sources, and all queries via global ontology. A domain expert, who understands the semantics of all data sources, needs to define the global ontology.	
Multiple ontology approach Each data source represented by its own local ontology. Mapping between the ontologies has to be established. No need to integrate data as changes to local ontologies may not affect the mapping.	
Hybrid ontology approach Data in each source are represented by a local ontology for which a shared vocabulary is built - intended to take advantage of the single and multiple approaches by simplifying local ontology definitions and querying via a shared vocabulary.	

Ultimately, the use/implementation of ontologies is highly dependent upon the healthcare setting/domain in which they are to be utilised. Research approaches to ontology matching may consider artificial intelligence, information system and databases perspectives. [133] Questions such as does an ontology already exist or require creation; does the data lend itself to a common vocabulary; what is the expertise and availability of key staff; what is the technology perspective; time constraints; and most importantly, the appropriateness of ontologies for the task. Hence, strategic coordination and investment in both human and technical resources are required to achieve semantic interoperability. [156]

2.4.2 Health Level 7

As patients move around the healthcare ecosystem, their electronic medical records must be readily available, discoverable, and understandable. In addition, data held within EMRs must be structured and standardised to support clinical decisions. [160] Therefore, the use of standards has become prerequisite throughout various medical systems. [161]

Healthcare data management is the domain in which the level of knowledge and proposed solutions to interoperability have accrued via years of research. [161] For over 20 years, Health Level 7 (HL7) has sought to address interoperability challenges by producing healthcare data exchange and information modelling standards. [160] Thus, within the field of healthcare information technology, HL7 is broadly acknowledged as the international standard for interoperability between health information systems. [23]

HL7 is an International Standards Development Organisation that produces the most widely used standards for healthcare interoperability in the world. The name HL7 derives from the application layer, which is the seventh level of the ISO's Open Systems Interconnect model that provides a framework for communication between disparate computer systems. [132] The seven layers of HL7 (adapted from Benson et al. (2016) [132] are presented in Table 2.16 (page 91).

Table 2.16 HL7 layers (adapted from Benson et al., 2016)

Area	HL7 layers		Description
Interworking	7	Application	Definitions of the data to be exchanged. Timing of interchange. Communication of any errors to the application.
	6	Presentation	Concerns the syntax of information transfer between end systems.
	5	Session	Mapping between physical and logical sessions. Includes checkpoint, recovery and restart.
Interconnection	4	Transport	End-to-end data transmission, e.g. free from error.
	3	Network	Routing and relaying between multiple sub-networks.
	2	Data-link	Transmitting stream of bits from one network node to another with error indication and limited error correction.
	1	Physical	Provide the interface to a physical communications medium.

However, HL7 does not develop software; it provides healthcare organisations with specifications to make their systems interoperable by developing coherent extensible standards using a formal methodology. This facilitates the exchange, management and integration of electronic healthcare information for both clinical and administrative purposes. [132]

HL7 Version 3 Normative Edition permits the exchange of messages based on a model-driven methodology that produces messages and electronic documents expressed in Extensible Mark-up Language (XML) syntax. [162] For example, a message is a string of textual information enclosed within tags, wrapped in angle brackets i.e. Start tags <tag> and End tags </tag>, and tags can may contain attributes i.e. <tag attribute="value">.

```
local XML_Template = [[
<?xml version="1.0"?>
<patients>
    <patient id="">
        <first-name></first-name>
        <last-name></last-name>
        <birth-date></birth-date>
    </patient>
</patients>
]]
```

Figure 2.12 Example of XML with patient details

Moreover, following technological advancements and development within industrial approaches and requirements, HL7 FHIR has emerged as a new specification. This next-generation standards framework created by HL7, known as Fast Healthcare Interoperability Resources (FHIR), and is designed to facilitate the exchange of healthcare-related information, e.g. clinical data, healthcare-related administrative, public health and research data. [160]

As FHIR focuses primarily on implementation, much of its specification targets organisations that are developing software and designing interoperable solutions to address the technical foundations of clinical information exchange between electronic systems. FHIR is applicable in a wide variety of contexts globally, e.g. in-patient, ambulatory care, acute care, long-term care, community care, allied health, and covers both human and veterinary medicine. [160]

The ‘Resources’ within FHIR define a small amount of highly-focused data and are an instance-level representation of a healthcare entity. Essentially, they are the common building blocks for all exchanges and represent the different types of clinical and administrative information that can be captured/shared. All ‘Resources’ possess common features such as a URL identifying the resource; common metadata; human-readable XHTML summary; set of defined data elements and an extensibility framework to support healthcare variation. [160] Within this context, metadata is defined as "data that describes data". [163]

Hence, a resource will only include data elements if there is an expectation that most implementations will use that particular data element. For example, the 'Resource Patient' contains detailed descriptions of elements such as Patient: identifier, name, birthdate, gender, marital status etc. [164] Table 2.17 (page 93) provides an example of the Resource: Patient, with Element: name.

Table 2.17 HL7 FHIR Detailed description of Resource: Patient, Element: name

Detail	**Resource: Patient Element 'name'**
Element Id	Patient.name
Definition	A name associated with the individual.
Cardinality	0..* *(i.e. one patient can have zero or many names).*
Type	HumanName.
Requirements	Need to be able to track the patient by multiple names. Examples are your official name and a partner name.
Summary	True.
Comments	A patient may have multiple names with different uses or applicable periods. For animals, the name is a "HumanName" in the sense that is assigned and used by humans and has the same patterns.

However, the capabilities of any FHIR-based system will be defined ultimately by what the 'Resources' can express from a clinical perspective. For example, the type of clinical record will dictate the kinds of Resources that are defined; their data contents; specific rules about the data, i.e. what terminology codes are supported and/or required; how resources reference each other and how to search for information. [160]

Moreover, HL7 FHIR facilitates straightforward implementation whilst ensuring the integrity of information, by leveraging existing logical/theoretical models to provide a consistent, implementable, and rigorous process for exchanging data between healthcare applications. [160]

Additionally, NHS Digital makes extensive use of this standard to support Application Program Interfaces (APIs), Documents and Messages, [165] thus, HL7 FHIR has fast become the leading open standard for interoperability. [161]

2.4.3 Knowledge discovery in databases

The term Knowledge Discovery in Databases (KDD) refers to the process of finding knowledge within data [166], the development of methods and techniques for making sense of data [167] via the application of specific Data Mining (DM) methods. [166] The KDD process addresses the main problem of mapping low-level data (typically too voluminous to understand and digest easily) into other configurations that may be more compact, abstract, or useful. [167] The authors Fayyad, Piatetsky-Shapiro and Smyth (1996), [168] define the term KDD as:

> "Knowledge Discovery in Databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". [168]
>
> - ***Data:*** *refers a set of facts, e.g. records within in a database.*
> - ***Pattern:*** *is an expression (in some language) describing a subset of the data or a model applicable to that subset.*
> - ***Process:*** *implies KDD is comprised of many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations.*

The KDD process is interactive/iterative and involves numerous steps that are essential to ensuring useful knowledge is derived from data-driven discovery. [167, 168]

1. Develop an understanding of application domain & prior knowledge to identify the objective.
2. Create a target dataset or use a subset of variables on which to perform discovery.
3. Data cleaning/preprocessing by accounting for missing data, noise, correct time-sequence.
4. Data reduction/projection via feature selection, dimensionality reduction, transformation.
5. Matching goals of KDD process to a data-mining method, e.g. classification, regression.
6. Exploratory analysis, select hypothesis, algorithms and methods to search for data patterns.
7. Data mining, search for patterns of interest, e.g. classification, trees, regression, and clustering.
8. Interpreting mined patterns, models, visualisations or redo steps 1 to 7 if required.
9. Acting on discovered knowledge, its direct use, incorporate into other systems, further actions.

Furthermore, KDD is interdisciplinary by nature and has evolved from the intersection of various research fields that include machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, data visualisation, and high-performance

computing. The unifying goal is to extract high-level knowledge from low-level data within the context of large data sets. [167]

Thus, the integration of medical information is achievable via the application of KDD techniques, [14] whilst data mining techniques facilitate the search for relationships and global patterns that exist within large databases but remain 'hidden' amongst vast amounts of data. [139]

Additionally, a common approach within the computing discipline for integrating data is to follow a specific sequence of steps, envisioned as a ‘data pipeline’, which typically refers to a chain of data processing stages, illustrated in Figure 2.13 (page 95), adapted from Burton et al. (2015). [169]

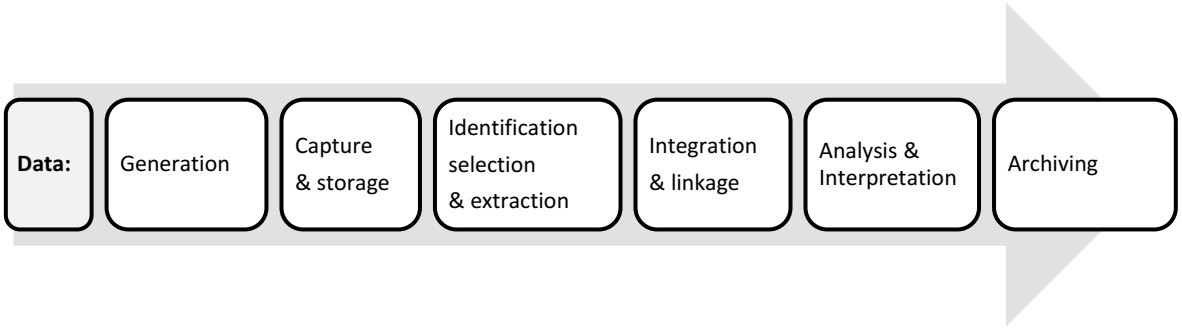


Figure 2.13 The data pipeline, adapted from Burton et al. (2015).

The data pipeline comprises two primary components that are temporally and/or spatially distinct, namely data acquisition and exploitation. Firstly, data acquisition subsumes the processes of data generation, capture, storage and archiving, whereby data originating in human/social context are collected within a data repository. Secondly, exploitation refers to the means by which these data are processed, managed, prepared and accessed, to enable health service provision; audit and evaluation; research use; analysis and interpretation. [169]

Consequently, a related field evolving from databases is ‘data warehousing’, which collectively refers to the process of collecting and cleaning data and making them readily available for subsequent analysis and decision support. Data warehousing helps to set the stage for KDD in two ways, first data cleaning and second, data access. [167]

First, data cleaning requires organisations to consider a unified logical view that encompasses the variety of data/databases they possess, whilst addressing the issues of semantic interoperability, i.e. mapping data to a single uniform representative naming convention, accounting for missing/noisy data and underlying inherent errors. Second, data access requires the specification of well-defined

methods to facilitate data access/retrieval taking into account syntactic interoperability issues, e.g. accessing historical data stored offline within legacy systems.

Moreover, database integration relates to the gathering and exchange of information from disparate and heterogeneous sources and requires bridging the syntactic and semantic gaps that exist amongst these sources. Thus, one of the first issues to consider is the location of the information to be gathered. [146]

2.4.4 Clinical data warehouse

Amongst heterogeneous hospital information systems, the integration of information at the data-level is inherently challenging, and to date, an integrated system which satisfies all the requirements of information processing within a hospital setting does not yet exist. [8]

However, the integration of disparate information from EMR and other healthcare information systems offers great potential for clinical care, public health practice, and research. [170] Hence, to improve/evaluate clinical practice as a whole, patient-specific information requires conversion into aggregate and multidimensional information. Such a conversion is made possible by employing data warehouse technologies within a healthcare setting to create a Clinical Data Warehouse (CDW). [114]

The CDW is not one thing per se [171], nor is it just a large collection of clinical data, [114] rather, it is a collection of components that encompasses different processes and houses a subset of data. [114, 171] It is characterised by its overarching purpose, to support organisations with specific retrospective analysis, decision-making and reporting tasks, [171] which may require specific architectural and technological consideration to optimise the CDW for aggregate analysis. [114] Hence, CDW's are not typically designed to be live or real-time systems. [114, 123]

Moreover, CDW's provide an information infrastructure, created by extracting and collecting clinical datasets from the EMR of multiple heterogeneous CIS into one central location. These datasets may be reorganised, filtered, coded, manipulated and linked, using common data elements to create an integrated dataset with common units and conforming dimensions. [114, 123]

CDWs are crucial to facilitate the management of an organisation's information, knowledge, and documentation. They enable managers, clinicians, analysts and researchers, quick and affordable means to obtain secure access; view detailed data held within patient records; instantly query information and undertake analysis to identify trends. [123, 136]

Furthermore, software applications used for running day-to-day business transactions, known as Online Transaction Processing systems (OLTP), incorporate functions geared to ensure 'live' transactions are handled in an accurately and efficiently. In contrast, the data-warehousing (DW) environment, also known as Online Analytical Processing (OLAP), seeks to support the decision-making processes by looking at data over time, often retrospectively, to identify any significant trends. [171]

Thus, from a healthcare perspective, OLTP is patient-focused, usually at the point of care. Clinical interventions are transactional in the sense that they are ordered, scheduled, performed, and documented within fairly well-defined clinical processes. However, the architectural design required for efficient OLTP is fairly inefficient for aggregate analysis and reporting. [114]

However, OLAP encompasses the conversion of transactional data into information, which can be used to guide and support organisations. OLAP has a population-based focus (i.e. patients attending a hospital), and facilitates retrospective or aggregate analysis; enabling clinical pathway development; intervention monitoring; evidence-based practice guidelines and evaluation of outcomes measures. [114]

The main differences between the OLTP and OLAP include functionality, performance, methodology, architecture, tools, and technology, Table 2.18. (page 98), adapted from Reeves (2013) [171] compares the key characteristics of the two technologies. However, OLAP technology is not designed to support operational systems; instead, it is intended to support applications in which read-only access to data is required. [116]

For example, when a patient presents to a hospital ED, an OLTP system, i.e. live departmental CIS, will be used to capture all details about their episode of attendance. Once the episode has concluded, and the patient has left the department, the recorded information is no longer 'live', hence becoming historical. Thus, all historical patient episodes can be accessed and viewed within the data warehouse or OLAP technology system for further analysis and reporting purposes.

Table 2.18 Comparison of online transaction processing and data warehouse systems, adapted from Reeves, 2013.

Characteristic	Online Transaction Processing (OLTP)	Data warehouse (OLAP)
Example	*Live Clinical Information System*	*Hospital Database Server*
System purpose	Support operational processes	Support strategic analysis, reporting
Data usage	Capture and maintain data	Exploit data
Data validation	Verification upon data entry	Verification after data entry
Update frequency	Live data updates	Updated via periodic schedules
Historical data	Current data	Multiple years of historical data
Data integration	Current system	Multiple systems

Hence, OLAP technologies are designed to facilitate data integration from multiple systems. Figure 2.14 (page 98) adapted from Reeves (2013) represents a basic data warehouse environment where four disparate data extracts are prepared, imported into the DW, organised and linked together, to facilitate access via users connecting to the DW for subsequent analysis. [171]

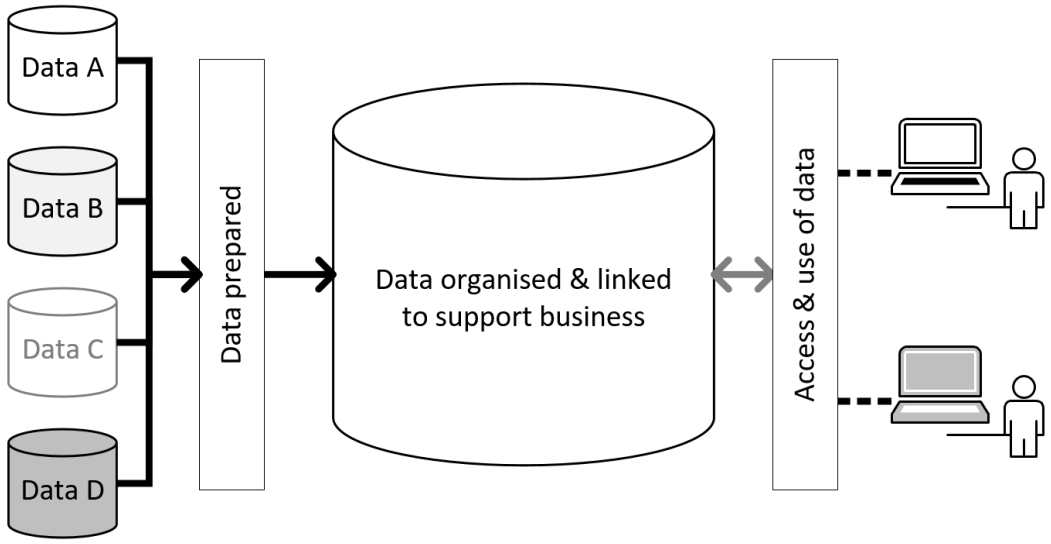


Figure 2.14 Basic data warehouse environment (adapted from Reeves, 2013).

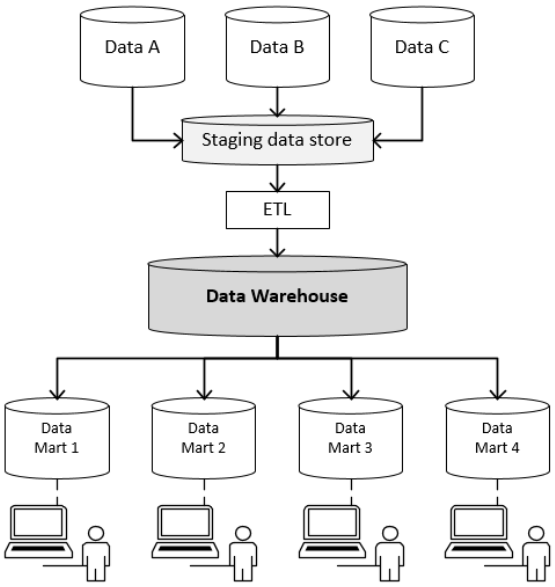
Nonetheless, when creating a CDW, one of the most important tasks is to gain a thorough understanding of the data itself. The author Reeves, (2013) [171] identifies preliminary factors that need to be considered, i.e. the type of data to be included; from which source systems is data being pulled; how the data has been organised; the level of detail required; the estimated size of the data; the need for data changes, coding, transformations; what level of summary is deemed appropriate and data provenance.

The authors Batini et al. (2006), [172] define data provenance as a "description of the origins of a piece of data and the process by which it arrived in the database." i.e., how and where it has been created, produced, and maintained. Typically, annotations are used to systematically trace the provenance and flow of data.

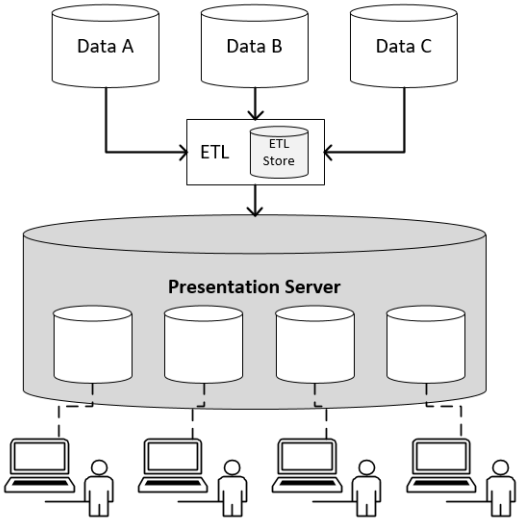
For example, within the field of data integration, gaining an in-depth understanding how data from different databases with heterogeneous semantics and different quality levels can be integrated, along with the processes to treat data, is crucial. [172]

Furthermore, when considering philosophical approaches to data warehouse architecture, the two most widely adopted are the 'Top-Down' approach and 'Bottom-Up' approach, as illustrated in Figure 2.15 (page 99), adapted from Reeves, (2013). [171]

Top-Down / Data-driven approach
Inmon DW architecture [173]



Bottom-Up / User requirements approach
Kimball DW architecture [174]



ETL: Extract, Transform Load

Figure 2.15 Comparison of Inmon and Kimball DW architectures (adapted from Reeves, 2013).

Top-Down, Inmon's approach [171, 173]

1. Capture/create the data: raw input data are extracted from various source systems, e.g. departmental databases ABC.
2. Extract the Data: extracts are stored within the Staging Data Store, where they are cleaned, validated, coded and transformed as required, additional supporting datasets, e.g. reference data may also be used.
3. Prepare the Data: datasets are then loaded into the Data Warehouse, where their structure is further assessed in relation to other datasets, considering semantic interoperability to facilitate data integration.
4. Publish the data: Data Marts are created to support specific business groups/areas to facilitate analysis, e.g. for a particular theme, inclusion/exclusion criteria are applied, and 'tailored' Data Marts created, to which users can then connect, to view and undertake data analysis.
5. Use the Data for Business Intelligence Applications, for example, creating tailored/bespoke Data Mart using integrated datasets, for a study looking at all hospital emergency admissions (via ED) within a specified time frame, where the patient is male, has Type 1 diabetes, is aged over sixty years, and resides within five miles of the hospital.

Bottom-Up, Kimball's approach [171, 174]

1. Capture/create the data: raw input data are extracted from various source systems, e.g. departmental databases ABC.
2. Extract the Data: involves gathering the data to feed to the ETL process.
3. Prepare the Data: the ETL Data Stores contain reference data used for validation, data quality checks and preparing dimensions, data are stored as simple flat-file structures, and only data required for next stage will be maintained.
4. Publish the Data: the Presentation Server delivers the 'core' data to the business user or analyst for querying, analysis, reporting, dashboard, and creation/use of multidimensional cube technologies.
5. Use the Data: Business Intelligence Applications, creation of robust multidimensional cubes to support a variety of reporting needs.

Ultimately, the specific business and systems requirements for each organisation will dictate which approach to use when developing DW architecture. [171]

Furthermore, there are two main techniques to data integration based on the actual physical location of data stored by various sources to be integrated. Firstly, virtual data integration, whereby

a unified 'view' of the data is virtual, with data residing only at sources. Secondly, materialised data integration, where the unified view of data is materialised and physically stored within the DW. [172]

DW offers a potential solution that allows data from multiple CIS to be homogenised and integrated within a data warehouse. Within healthcare, this technique has proved to be a useful way of bringing disparate data together in a single repository, thereby creating an integrated EMR. [116]

However, the benefit of obtaining data from multiple sources comes at an additional cost of building elaborate algorithms that hide the underlying complexity of maintaining multiple copies of a single data element. The fundamental difficulty lies in keeping the multiple copies consistent with each other, while at the same time maximising data availability and performance. [8]

In addition, integration at the semantic level is achievable by the transformation of data sources, which supports the resolving of heterogeneity conflicts and facilitates the matching/merging of individual data/schemas, resulting in a unified/global and integrated view. Processes such as 'data object fusion' seek to identify semantically equivalent data objects to determine a consistent representation amongst various datasets. For example, abstraction includes functions to transform source data via aggregation or reclassification; or supplementation via the addition of semantic metadata (e.g. in the absence of a schema), can facilitate the fusion process. [133]

Moreover, to aggregate and integrate health information, patient record linkage (matching) is an important process for identifying and connecting records belonging to the same patient across several disparate data sources and over extended periods. [175] Nevertheless, it is the actual data that put the 'evidence' into 'evidence-based medicine', which is crucial to supporting developments within Public Health, Primary and Secondary Care. [169]

2.4.5 Section summary

This section provided an overview of techniques that focus on data integration by describing the four types of system interrelationships and associated levels of communication. It then introduced ontologies (within the field of Information Systems) as explicit specifications of a conceptualisation, and how ontologies have the potential to address semantic interoperability. The healthcare interoperability standard HL7 was presented, along with the most recent specification HL7 FHIR,

targeted to addressing the technical foundations of clinical information exchange between electronic systems. Additionally, to address syntactic interoperability, an overview of Knowledge Discovery in Databases and the Data Pipeline was provided along with a summary of the Data Warehouse, on the whole, these techniques sought to integrate disparate medical information.

2.5 Data quality

NHS England describes information as its lifeblood, a vital resource for the clinical management of individual patients and the efficient management of services. [25] Within a hospital setting, increasing amounts of information are transcribed routinely within EMR by both administrative and clinical staff. [176] However, the manner in which the information is recorded is affected by the medical specialty; local clinical or administrative process; individual staff approaches; departmental policies; and ultimately the nuances and/or limitations of the operational host CIS. [177]

Although hospital patient data are routinely collected within CIS for the management and treatment of patients, such data is also used for medico-legal, clinical, financial and other purposes. Moreover, routinely collected data have the potential for re-use in many other tasks, including research. [178, 179]

Thus, from a research perspective, a valuable feature of routine electronic data is its availability, in that it may be relatively quick to obtain and may already be processed to certain specifications, and relatively lower in cost when compared to primary data collection. Therefore, the most straightforward approach to obtaining such data would be from the CIS in which the data are stored. [179-181]

However, the original objective/purpose for the collection for routine data will ultimately influence and constrain the type of research questions, their level of depth and type of analysis. [179] In addition, there some important factors to consider when accessing routine data, first, legal access, can potential users of data obtain access legally and ethically whilst conforming to all requirements of the providers and meeting information governance standards and establishing data ownership and guidelines to facilitate access. [181]

Second, the financial implications of accessing data, i.e. the cost to the organisation, in terms of resources required, availability of specialist/expert personnel, user training, and secure extraction

and storage requirements. [181] Therefore, advice and assistance from data custodians will need to be sought to ascertain what level of access is warranted, and the level of detail required by potential users. [180]

Nonetheless, there are potential caveats to the use of such data, as there is no control over what data are available and limited control over missing data or the ability to fill gaps retrospectively or resolve queries. In addition, data may not be in the required format, or 'raw' data may contain inconsistencies, be inaccurate, incomplete or transformed in ways that undermine their meaning, thereby rendering such data unsuitable for research. [177-179]

Moreover, information exists to support decisions and actions, and if it fails in this task, it could be considered irrelevant noise. [98] Thus, data that do not accurately represent the information they are intended to reflect, i.e. data that have been transposed; scrambled; omitted; are full of errors; incomplete or cannot be interpreted by the user, can be thought of as being corrupted and as such, unlikely to be of benefit. [98, 112]

Additionally, the heterogeneity amongst CIS will strongly influence the properties, formats, types and values of data contained within each CIS, and the provenance may by unclear, data may also be of insufficient granularity and incompatible with research protocols. [177-179] Ultimately, the quality of the underlying data will influence the strength of any type of research or evaluation. [93]

Figure 2.16 (page 103), adapted from Iyamu et al. (2018), [182] illustrates the key challenges that exist when considering healthcare data analysis, e.g. data storage, processing, analysis, inconsistency, validation and quality. The author reports such challenges are linked directly to the specific healthcare settings and activities for which their data are recorded.

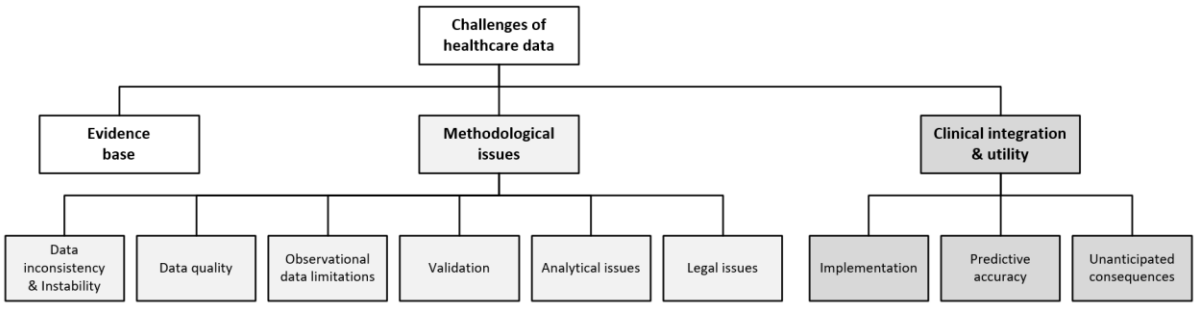


Figure 2.16 Challenges for healthcare data (adapted from Iyamu et al. (2018)

2.5.1 Data quality as 'fit for purpose'

At present, there is no universal agreement of how best to describe Data Quality (DQ) within the research domain. [163] However, The International Standards Organisation (ISO), within the domain of 'Systems and software engineering' [183] define DQ as: "The degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions".

Correspondingly, an article by de Lusignan (2006), [184] identified one definition of DQ as 'fit for purpose', which implied DQ dimensions such as accuracy, completeness, relevance and accessibility "were not ends in themselves" but more importantly, "whether data were suitable for their intended purpose". The author concluded by proposing a more functional approach, i.e. "are these data fit for the purpose intended?". [184]

Additionally, DQ encompasses the specific characteristics of data conveyed through its metadata, i.e. additional information that describes/assesses the data. [185] A review of the literature by the authors Liaw et al. (2013), [186] acknowledged that although there was no consensus as to the definition of 'DQ dimensions', the five most commonly reported dimensions were accuracy, completeness, consistency, correctness and timeliness.

Therefore, a DQ dimension is defined as 'a set of DQ attributes that represent a single aspect or construct of DQ'. [187] Whilst the term 'DQ', encompasses several dimensions and is considered to be the degree to which data items are accurate, complete, relevant, timely, sufficiently detailed, and appropriately represented whilst retaining sufficient contextual information to support the decision-making process.[26]

The authors Wang et al. (1996), [187] identify three approaches to study DQ within the literature. First, the Intuitive approach, in which the selection of DQ attributes for any particular study is based on the researchers' experience/intuitive understanding of what attributes are considered important. Second, the Theoretical approach to DQ, which focuses on how data may become deficient during the data manufacturing process.

Third, the Empirical approach to DQ, which analyses data from the perspective of the data-consumer to determine the characteristics they use to assess whether data are 'fit for purpose'. However, although the concept of 'fit for purpose' is widely adopted in the literature, it emphasises the

importance of taking a consumer viewpoint of quality, because ultimately the consumer will decide whether something is fit for use. [187]

Figure 2.17 (page 105) adapted from Wang et al. (1996), [187] illustrates a conceptual framework of DQ.

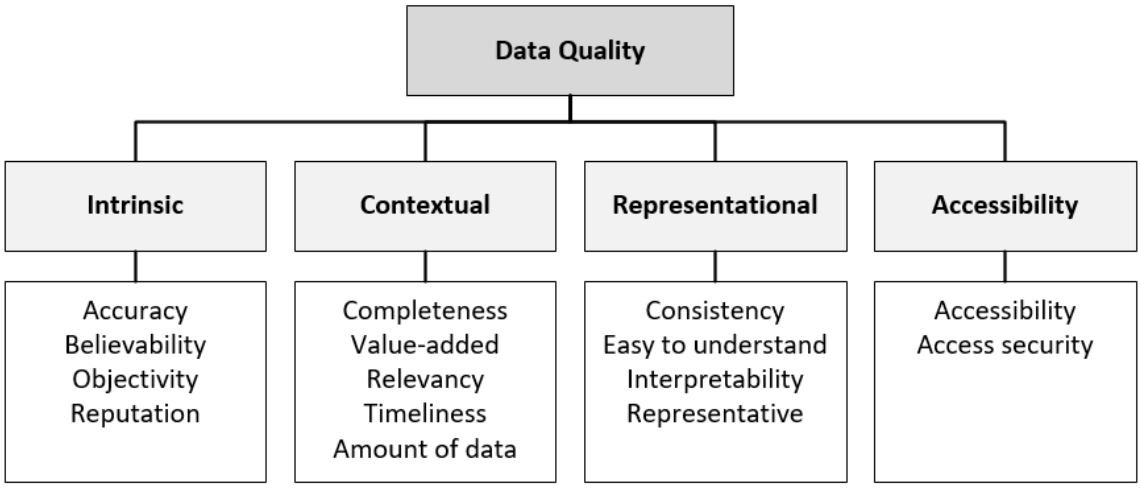


Figure 2.17 Conceptual framework of data quality (adapted from Wang et al., 1996)

Moreover, data is not just collected unintentionally; it has a purpose to exist and therefore, must be utilised in a manner in which it fulfils that purpose. Thus, data can be considered to 'have quality' if it satisfies the requirements of its intended use, and to achieve this data must be trusted and understood. Conversely, data can be considered to 'lack quality' to the extent that it does not satisfy that requirement. Nevertheless, DQ depends as much on the intended use as it does on the data itself. [188]

Table 2.19 (page 106), provides an overview of key DQ dimensions: completeness, uniqueness, validity, accuracy, consistency and timeliness, with associated descriptions and examples of calculations.

Table 2.19 Data quality dimensions and associated calculations

Dimension	Detail
Completeness	○ The proportion of stored data against the potential of '100% complete' ○ Business rules which define what '100% complete' represents ○ Measure of blank (null/empty) values or the presence of non-blank values ➔ Unit: Percentage ○ E.g. Total patients 900 records, only 786 have postcode recorded ⇒ (786÷900) ×100 = 87.3% complete
Uniqueness	○ No data item recorded more than once, based upon how item identified ○ Data item measured against itself or counterpart in another dataset ○ Number of data items as assessed in the 'real world' compared to the dataset ➔ Unit: Percentage ○ E.g. Total 525 unique postcodes, total record is 786 postcodes ⇒ (525÷786) ×100 = 66.8% uniqueness or duplicate records ⇒ (786÷525) = 1.5 level of duplication
Validity	○ Valid if data conforms to the syntax (format, type, range) of its definition ○ Refer to metadata rules for allowable types, e.g. string, integer, date, time ○ Format, e.g. length, number of digits ○ Range, e.g. minimum, maximum, or within a set of allowable values ○ Comparison between the data items and metadata rules ➔ Unit: Percentage of data items deemed valid or invalid ○ E.g. 20 invalid postcodes within 786 postcodes ⇒ (20÷786) ×100 = 2.5% Invalid postcodes ⇒ (766÷786) ×100 = 97.5% Valid postcodes
Accuracy	○ The degree to which data correctly describe the 'real world' object or event ○ Reference 'real world' via primary research, or equivalent related source ○ Scope 'real world' objects characterised by data item ➔ Unit: percentage of data entries that pass the data accuracy rules ○ E.g. Date of birth UK DD/MM/YYYY, but recorded MM/DD/YYYY ⇒ $\frac{(Count\ of\ accurate\ objects)}{(Count\ of\ accurate + inaccurate\ objects)} \times 100$
Consistency	○ No differences when comparing data item representations against a definition ○ Data item measured against itself or counterpart in another dataset ○ Analysis of pattern and/or value frequency ➔ Unit: Percentage ○ Assessment of data items across multiple data sets ○ E.g. Patient ID is the same value and format within two hospital systems ⇒ Select count distinct on 'Patient ID'
Timeliness	○ The degree to which data represent reality from the required point in time ○ Refer to time the real-world event being recorded occurred ○ Usually affected by the way data are collected ○ Time acts differently on static and dynamic record ➔ Unit: Time ○ E.g. A clinician ended a consultation at 1645 and completed a paper record. Later at 1700 the clinician retrospectively completed an electronic transcription of the record ⇒ 1700 – 1645 = 15 minutes time difference.

Content adapted from DAMA UK Working Group [189] and CDC [190]

Detailed definitions of DQ dimensions have been extensively covered within the literature, including the following sources [98, 181, 186, 188-191], but not limited to these. However, brief descriptions of the DQ dimensions listed in Table 2.19 (page 106) are provided henceforth.

The dimension 'Accuracy' refers to whether the data values stored for an object are the correct values [188], i.e. if the recorded value conforms with the actual value, [186] and the extent to which the data are free of identifiable errors. [181] In addition, accuracy can be tested by comparing the specific data value with a gold standard or a rule, e.g. ensuring the height of a patient cannot be recorded as five inches. [98, 181]

Furthermore, checking for value 'Consistency' and representation, e.g. Street vs St or the identification of an unambiguous value 'Unknown' within a Name data element. Missing values may be left as 'blank' cells, or in other cases encoded as 'NULL'. Inconsistencies may also exist within data tables, e.g. the unintentional inclusion of 'Test' patient records when undertaking analysis. [181, 188]

The issue of 'Timeliness' may persist if a patient's EMR was entered retrospectively and the departure time recorded before the arrival time. Structural analysis techniques involve checking columns of values for uniqueness, e.g., two patients with different demographics being allocated the same patient number. [181, 188]

Likewise, another approach to identifying an inaccurate value is to use 're-verification', in which each data value is checked manually against original information. However, this process is time-consuming and impractical. A more robust approach is to use analytical techniques, which begin with a definition of what is deemed 'correct' for each data element by looking at the attributes of individual values in isolation to determine if they are valid, e.g. examine data type, length, range of values, list of discrete values allowed. [98, 188]

Nevertheless, the process of assessing DQ dimensions is achievable via careful planning and management and involves setting well-defined DQ goals and standards, along with definitions of data collection strategies to undertake an assessment of collected data using DQ indicators. [185, 192]

2.5.2 NHS number validation

The NHS number is the primary/unique identifier allocated to patients registered for NHS care in England, Wales or the Isle of Man. It is valid for life and assigned to all newborns or patients the first time they receive NHS care/treatment and enables healthcare service providers to correctly identify patients and match their details to their respective health records. [193]

In addition, the NHS number consists of 10 numeric digits in length, and to ensure consistency and prevent errors, a five-step algorithm illustrated in Figure 2.18 (page 108) and detailed in NHS Data Dictionary [194] is utilised for validation purposes. For example, take a spurious number: 1234567890. The last digit on the right-hand side (check digit) is removed and each individual remaining number multiplied by its position in the sequence to calculate the factor, which is then summed to obtain the total. Thereafter, the total is divided by 11, and the remainder (modulus) obtained to one decimal place. Finally, if the result of the calculation 11 minus the modulus is equal to 10, then the NHS number is invalid.

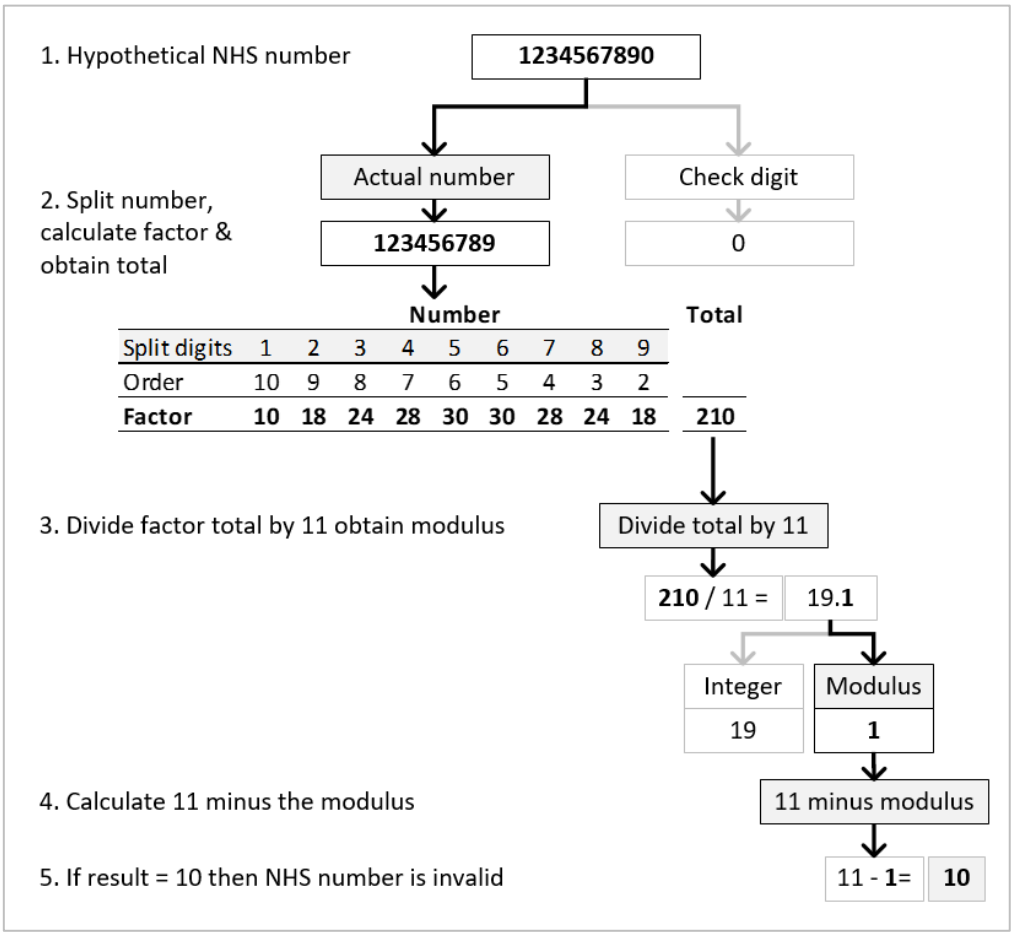


Figure 2.18 Validation process for the NHS number

2.5.3 Data quality management

When it comes to the matter of extracting routinely collected healthcare data, e.g. EMRs from a hospital CIS, researchers may know very little about what to expect regarding the level of DQ for such extracts. In addition, as routinely collected healthcare data are being used increasingly for analytics, quality improvement or research, they will require data extraction processes that are deemed fit for purpose. [195]

Furthermore, effective assessment and management of DQ requires analysis of the entire data cycle, i.e. from collection through extraction, cleansing, storage, management, dissemination, presentation and curation. [163, 195] Figure 2.19 (page 109) adapted from Davoudi et al. (2015), [181], provides a simplified overview of various DQ management challenges relating to the integrity of data, that will need to be considered by anyone wishing to undertake research.

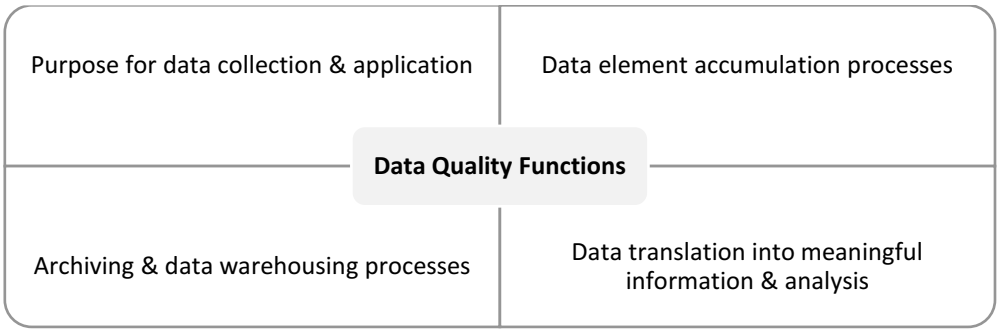


Figure 2.19 Data quality functions (adapted from Davoudi et al., 2015)

First, considering the purpose for data collection, its application and plan of analysis in reference to the research question. Data collection should be timely and limitations well documented, to ensure data are as complete as possible. For instances of data extraction from multiple applications or systems, the 'value' (meaning) of the data should be semantically identical, and subsequent analysis should align with and reflect the research question. Additionally, researchers should ensure only specific data elements required for the research are selected.

Second, concerning the process by which data elements are to be collected/accumulated. Data definitions should be clear and concise, aligned to appropriate data standards, along with descriptions of acceptable values/ranges for individual data elements, and accessible to researchers, e.g. data dictionaries. The data collection process should be standardised where possible, and the collection instrument validated, e.g. data distortion by extraction/management tools employed. In

addition, data sources should provide the most accurate, timely, and least costly data whilst providing an appropriate level of detail/granularity for analysis. Furthermore, researchers should undertake any training as appropriate to gain a better level of understanding and insight into the provenance of data and to support the research process.

Third, relating to data warehousing, archiving and interoperability. Any data imported into the data warehouse should possess an appropriate level of granularity/detail to facilitate analysis and answer the research question. Regarding data ownership, relationships between the roles of data owners, collectors and end-users will need to be defined, in light of information security and governance policies for access to data/systems. Additionally, a data inventory, retention schedules and guidelines will support archiving and maintenance tasks, in accordance with research protocols.

Furthermore, the incorporation of bespoke conversion/translation tables and data definitions within the DW will facilitate semantic interoperability between datasets originating from different sources, further supported by relational schemas to provide researchers with a 'roadmap' to understating any merged datasets.

Fourth, regarding the translation of data into meaningful information by undertaking analysis. Following the incorporation of multiple datasets from different sources, with appropriate granularity, and referential integrity, the creation of new linked/merged datasets was attainable within the DW. However, DW technologies have limited statistical analytical capability; therefore, practically all DW technologies facilitate connectivity to other applications. Hence, data remain within the DW, and a secure connection to external analytical tools, e.g. Statistical or Business Intelligence Applications facilitates data access and analysis. Thus, researchers are able to connect their preferred analytical tool to the DW in order to undertake any analyses.

Likewise, the authors Liaw et al. (2014), [192] identify six preliminary questions relating to DQ management that researcher (or anyone working with data) should consider:

1. Why am I collecting/recording these data? (Data collection and utility)
2. How did data come to be? (Data provenance including metadata)
3. Do data look right? (Errors with data extraction, linkage, processing and translation)
4. Who or what can I check the accuracy with? (Triangulate and validate data iteratively)
5. Where did data come from? (Traceability)
6. How do I look up earlier data? (Curation)

To conclude, routinely collected healthcare data within EMRs are increasingly being mined and aggregated into data repositories or CDW, linked and used for audit, continuous quality improvement in clinical care, health service planning, epidemiological study and evaluation research. [196] Such data used sensibly and in conjunction with other data sources can add significant value to any research project. [176]

However, the reuse of routine data to advance clinical research can be challenging, as the quality, timing, and comprehensiveness of such data may not be consistent with research standards, [178] which may increase the likelihood and scope of data errors. [196] Likewise, poor quality data may lead to a lack of clinical confidence, mistrust, increase the potential for clinical error or incorrect decisions, and invalidate any analysis undertaken. [176]

Moreover, the reliability and applicability of results following data analysis are dependent on the quality of the original data providing the background for the assessment. [185] Not only is the quality of data important, so too is its management, as poorly managed data is subject to hidden issues that may distort the original dataset and therefore its value in reaching reliable conclusions.

Nonetheless, the quantity and real-world nature of routine data provide a strong impetus for their use. [178] Hence, to effectively assess and manage DQ, the entire data cycle from collection to extraction, data cleansing, storage, management, dissemination, presentation and curation, will require meticulous consideration. [163, 195]

Although DQ is a multidimensional concept lacking consensus framework and definitions, [186] the question ‘are these data fit for the purpose intended?’ [184] may help contextualise it. Thus, to improve DQ, an understanding of what it means to data consumers (those who use data) is required, as ultimately they decide whether something is fit for purpose. [187]

2.5.4 Section summary

This section discussed the increasing use of routinely collected hospital data for research and decision-making purposes. It emphasised that the objective/purpose for data collection should determine what type of data were to be collected, as this would subsequently restrict the types of research questions being asked and analysis undertaken. Additionally, the caveats of using routinely collected data within the context of data extracted from multiple heterogeneous information

systems were highlighted. The concept of data quality as 'fit for purpose' and an overview of various DQ dimensions, along with definitions and calculations provided. Equally, the importance of considering DQ management techniques and preliminary questions researchers should contemplate/address was emphasised.

2.6 Data linkage

The integration of disparate information contained within EMRs across different healthcare systems and clinical specialties within the NHS offers great potential for clinical care, public health practice, and research. However, such integration depends on correctly matching patient-specific records using demographic identifiers. Due to the pervasiveness of structural and semantic data heterogeneity, and without acknowledged standards for personal identity attributes or identifiers, the process of data linkage is intrinsically complicated. [170]

Additionally, semantic heterogeneity arises from the differences in the meaning/content of demographic identity fields in disparate information systems, whilst structural heterogeneity occurs due to variation in the levels of quality, completeness, and formats for storing identifying information across different information systems. [170]

Thus, In 1946, Dr.Halbert Dunn, [1] first described a book of life, that covered an individual's record of events from cradle-to-grave. The process of assembling the pages of this book into a volume was described as 'Record Linkage'. Dunn also emphasised the importance of record linkage to both the individual and to health, welfare and other organisations.

The terms 'Record linkage' (RL) and 'Data linkage' are often used interchangeably within the literature, and may also be referred to as 'Entity resolution' or 'Object identification' by computer scientists. [197] Therefore, RL can be conceptualised, as the process of searching for records that belong to the same individual [198] from two or more distinct sources and bringing that information together. [199]

Furthermore, RL is a powerful tool that can be used to identify multiple appearances of individuals within a dataset, enabling deduplication of individual records, facilitating the integration of information across datasets and building longitudinal profiles. [199]

Routinely collected administrative data comprise information already widely and diligently collected for specific purposes. Such data have the potential to be linked at an individual level, within administrative databases, using unique identifiers that do not compromise confidentiality. Hence, RL techniques are widely applied by many organisations, e.g. businesses, government, and research institutions. [197, 200]

Advancements in research and technology have enabled computerised RL techniques to make use of existing datasets, enabling information to be combined efficiently and cost-effectively, avoiding the high cost, time and effort associated with setting up new research studies, [201] while additionally resulting in substantial improvements of data quality. [197]

Moreover, RL is bringing together information from two records believed to relate to the same entity (e.g. an individual). This may involve linking records within a single database to identify duplicate case records or involve linking records from two or more databases, to merge them. [197]

For example, RL is straightforward when a unique identification number, e.g. Social Security Number or NHS number, is readily available. However, it becomes more challenging when only quasi-identifiers, i.e. an individual's demographic information (forename, surname, date of birth or address) are available. [202] One potential strategy employed to address such instances, is to use a combination of quasi-identifiers to enable the 'unique' identification of an individual. [197]

However, information recorded for an individual may vary from dataset to dataset, due to either difference in reporting (e.g. in the first name) or inherent errors within the transcription process; hence, linkage processes should allow for some discrepancy in reported characteristics. [203] Nevertheless, the primary objective of RL is to find matches [198], and the two predominant approaches within the research literature are Deterministic RL and Probabilistic RL. [204]

2.6.1 Deterministic Record Linkage

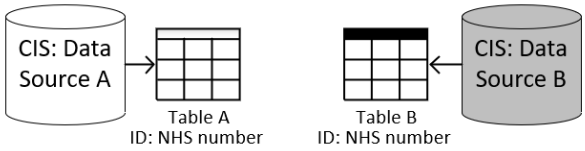
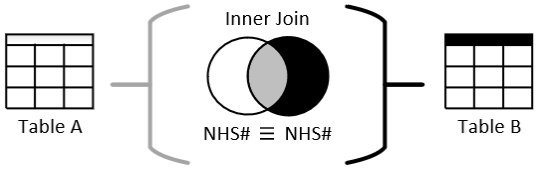
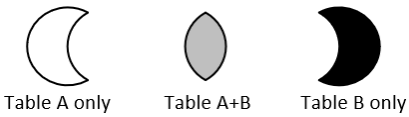
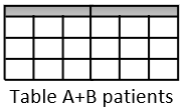
Deterministic record linkage (DRL) is the linkage of two (or more) files based on an exact agreement of matching data elements. [198, 202] To illustrate the concept of DRL elements that match exactly, an example using the NHS number is provided in Table 2.20 (page 114). The NHS number is the primary and unique identifier for a patient within the NHS in England and Wales. [193]

First, specific data elements are selected from two sources CIS A and CIS B and extracted into two distinct tables A and B, respectively. Each table contains records about the patient episode of care, and each patient record has been allocated has a unique NHS number.

Second, the objective is to identify which patient records appear within both tables A and B, achieved by linking on the identifier NHS number (from each table) to obtain an exact match. Third, the Venn diagram depicts three distinct segments, i.e. patients only in Table A, patients only in Table B and the intersection clearly identifies those patients appearing in both tables A and B.

Therefore, when a comparison is made between two datasets, the unique identifier common to both databases becomes the matching data element, used for linking records, [198, 205] which from this example, is the NHS number. DRL is the process of linking two files together using, either one data element (e.g. unique ID) or multiple data elements that match exactly.

Table 2.20 Example of deterministic record linkage

#	Process	Representation
1.	Data extracted from two CISs as two distinct Tables A and B, containing details about the patient episode of care. Each patient has a unique NHS number. Both tables are assessed for DQ.	
2.	The Venn diagram intersection (grey) represents an 'Inner Join' where there is an exact match between the NHS numbers from both tables.	
3.	Three distinct queries for patient records are possible. Those appearing only in Table A, or Table B or those in linked Table A+B.	
4.	The result of the intersection is a linked table containing patients that appear in both Table A and B.	

2.6.2 DRL strategy: Hospital Episode Statistics

Within NHS England, the principal source of patient-level hospital data is the Hospital Episode Statistics (HES) database which is a repository of information about all admissions, ED attendances and outpatient appointments, as recorded by the provider.[206, 207]

Following a patient's treatment at a hospital, the collected information forms part of the Commissioning Data Set (CDS) which is submitted to NHS Digital for processing and made available to commissioners and healthcare providers of NHS-funded care as the 'Secondary Uses Service' (SUS). The term 'Secondary Uses' indicates its utilisation for non-clinical purposes, i.e., not relating to direct patient care. Thereafter, SUS data are extracted into the HES dataset, Figure 2.20 (page 115) provides an illustration of the overview of the data pathway from CDS to HES.

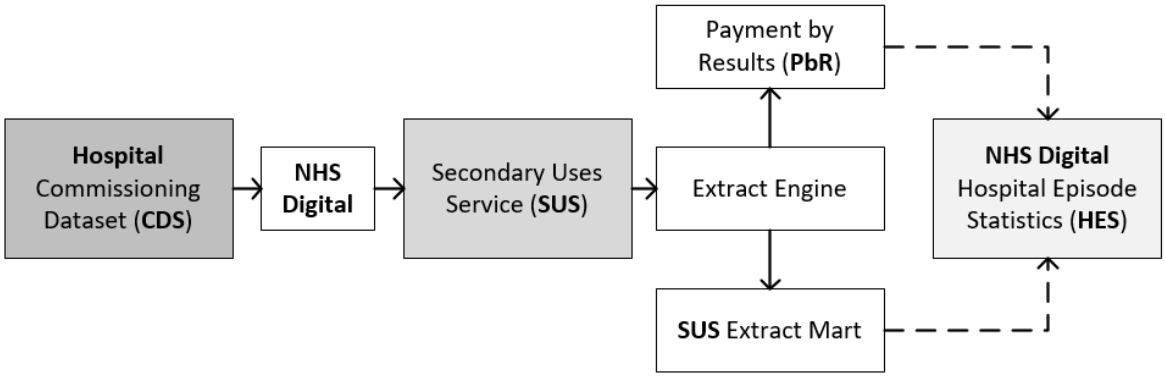


Figure 2.20 Path of hospital data extracts into the HES database.

Furthermore, the HES dataset is used for Healthcare planning, Commissioning of services, National Tariff reimbursement, Development of national policy and Research. HES contains data for all admitted patients since 1989, ED data since 2007 and outpatient data since 2003. Every year the number of patient data increases by an estimated 19 million inpatient records, 18 million ED records and 90 million outpatient records. [206, 207]

Thus, HES is effectively a secure data warehouse that stores patient-level information e.g.

- Administrative information, e.g. appointment dates, admission method.
- Patient demographic/biographic information, e.g. age, gender, ethnicity.
- Clinical information, e.g. diagnoses and operations.

The HES dataset contains personal details that help to identify individuals, e.g. NHS Number, Date of Birth, Gender and Postcode. When linking datasets, these patient identifiers are matched using a

'Step-based DRL' strategy [206] based on combinations of the NHS number and aforementioned quasi-identifiers, as described in Table 2.21 (page 116).

Table 2.21 Step based approach for HES matching criteria

Step	NHS number	Date of birth	Gender	Postcode
1a	Exact	Exact	Exact	Exact
1b	Exact	Exact		Exact
2	Exact	Exact	Exact	
3	Exact	*Partial*	Exact	Exact
4	Exact	*Partial*	Exact	
5	Exact			Exact
6		Exact	Exact	Exact
7		Exact	Exact	Exact
8	Exact			

Moreover, this step-based approach seeks to identify potential matches between two datasets, e.g. Step 1a requiring strict, exact match criteria on all four data elements and Step 1b requiring exact match on three data elements. Additional criteria for Step 6 and Step 7, stipulate the date of birth cannot be 1st January, and the postcode cannot be from a communal establishment, e.g. hospital, prisons, army barracks. Whereas, Step 8 is more relaxed than the previous steps, but requires clerical review. Thus, following each step in sequence effectively reduces the remaining number of records for the next step. [206]

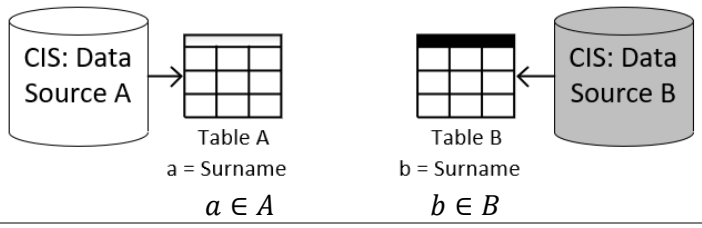
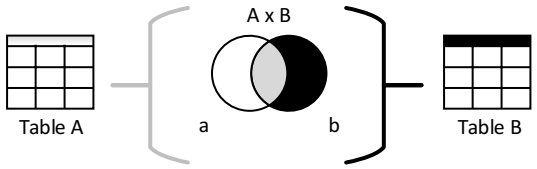
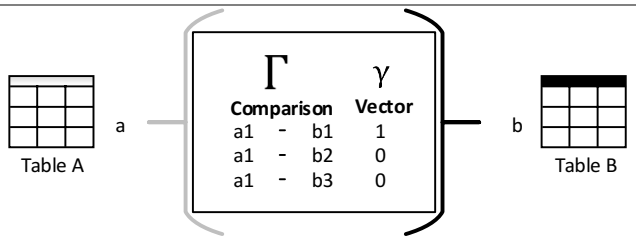
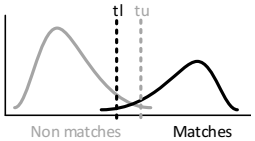
2.6.3 Probabilistic Record Linkage

Howard Newcombe's pioneering, practical work on computerised RL began in the 1950s and culminated in an important article 'Automatic Linkage of Vital Records'. [201] Newcombe, a geneticist motivated by his interest in genetic and biomedical research recognised that RL was a statistical problem, i.e. when comparing a pair of records using 'identifying information' that contained potential errors, how to decide best which pair of record comparisons should be regarded as truly linked. [208]

Thus, in 1969 Fellegi & Sunter [209] developed a theory for RL and formalised a mathematical model to incorporate probabilities, based on the ideas introduced by Newcombe. [201] Universally known as 'Probabilistic Record Linkage' (PRL), the model involves the linkage of two (or more) files, calculation of weights for agreeing and disagreeing data element pairs, deriving a match weight (represents the likelihood of records belonging to the same individual) and classifying record pairs to

enable optimal decision-making. [198, 209] To illustrate the concept of PRL and the Fellegi & Sunter model, Table 2.22 (page 117) presents a high-level example.

Table 2.22 Simplified overview of Fellegi & Sunter PRL model.

#	Process	Representation of theory
1.	Data extracts from CIS A and B Two distinct Tables A and B e.g. Data item Surname from each table, where a is an element of A b is an element of B	
2.	Product space $A \times B$ results in a Cartesian product (all possible record pair comparisons) A blocking field is selected, e.g. joining on surname initial	 $\mathbf{A} \times \mathbf{B} = \{(a, b);\ a \in \mathbf{A}, b \in \mathbf{B}\}$
3.	Two disjoint sets:	Set M: Matches $M = \{(a, b); a = b, a \in \mathbf{A}, b \in \mathbf{B}\}$ Set U: Non-matches $U = \{(a, b); a \neq b, a \in \mathbf{A}, b \in \mathbf{B}\}$
4.	Within product comparison space Γ (Gamma) Calculate agreement pattern Vector γ for each record pair → $\gamma \in \Gamma$ Vector γ is binary (1, 0)	
5.	Calculate ratios (r) conditional probabilities from the comparison of a x b	Conditional probabilities $R = \frac{P(\gamma \in \Gamma \mid r \in M)}{P(\gamma \in \Gamma \mid r \in U)}$ b: Match / Non-Match a: Link (1 1 / 1 0), Non-Link (0 1 / 0 0)
6.	Calculate weights for data element	Match if a = b $w = log_2\left(\frac{m}{u}\right)$ Non match if a ≠ b $w = log_2\left(\frac{1-m}{1-u}\right)$
7.	Summing all the weights for all data elements provides the total weight (w_i) (inverse of the total weight is probability)	Γ: Comparison (a1 - b1; *ai - bi*) · γ: Vector (1; *1 or 0*) · Probability (M; *M or U*) · Weight (w1; *wi*) $log_2(R) = \sum w_i$
8.	Obtain weight distribution & calculate thresholds $t_u = Upper\ threshold$ $t_l = Lower\ threshold$	 $R \geq t_u \Rightarrow r \rightarrow$ Match, $t_l < R < t_u \Rightarrow r \rightarrow$ Potential Match, $R \leq t_l \Rightarrow r \rightarrow$ Non-Match.

Step one: Data were extracted from two hypothetical CIS (A and B) and stored as two distinct data Tables A and B. Each table contains details about patient episodes of care. However, there is no unique patient identifier (ID); therefore, DRL techniques cannot be used. Hence, the objective is to identify which patients appear in both tables using patient quasi-identifiers, i.e. surname, forename, sex, date of birth and postcode, within each table.

Step two: The comparison space $A \times B$ results in a full Cartesian product of ALL possible record pairs generated from the two databases being matched. [177] For example, comparing the surname 'Black' from Table A, with every other surname within Table B would result in many unnecessary multiple 'comparison pairs', such as Black vs May which clearly are not the same. A comparison pair is any possible comparison of a record from one file with a record from another file. [198]

Therefore, to reduce the number of comparisons, the first letter of the patient's surname, i.e. the initial, is selected as the 'blocking' (or partitioning) data element, common to both tables A and B. Thus, the comparison of record pairs is restricted to only those within these blocks, ultimately confining the surname Black to comparisons with other surnames only beginning with the letter B. Moreover, blocking techniques greatly reduce the number of comparisons and improve efficiency. [198] Ideally, the actual values within blocking fields (data elements) should possess high levels of variety, and a low rate of missing values. [210]

Step three: Matches between the two data sets are then restricted to comparison pairs within specific blocks. [211] As defined by Blakely et al. (2002), [198]:

- A **Match (M)** is defined as A comparison pair of records that are for the same person.
- A N**on-match (U)** is defined as A comparison pair of records that are not for the same person.

For example, if the surname initial agrees and the surname agrees, i.e. B ≡ B and Black ≡ Black then the M probability can be calculated, i.e. matching data elements agree, given that the comparison pair being examined is a match. [198] Additionally, if the surname was miscoded in 5% of record pair comparisons, the m-probability would be 0.95. [204]

Conversely, if surname initials B ≡ B but actual surnames Black ≠ Brown, this is the U probability, i.e. the matching data element agrees given that the comparison pair being examined is a non-match, which indicates the probability that the data elements agree purely by chance. [198] For example,

the probability of chance agreement on sex is 1/2 (Male or Female) similarly for the month of birth is 1/12 (as twelve months in a calendar year).

Step four: The agreement patterns for all supplementary data elements are then calculated, i.e. for comparison pairs where every surname initial matches (B ≡ B), additional check are undertaken to verify whether or not details from Table A (surname, forename, date of birth, sex, postcode) exactly match Table B (surname, forename, date of birth, sex, postcode).

The result of this comparison is termed the comparison vector (γ) and the outcome for each vector is binary, i.e. IF Black ≡ Black Then 1, Female ≡ Female Then 1, Postcode ≠ Postcode Then 0. Therefore, a key assumption within the Fellegi & Sunter formulation is one of conditional independence between the data elements. [199]

Step five: Ratios are then calculated by looking for exact agreement or disagreement between matching data elements between files. As the true match status is unknown, a link status is established. Whereby, a 'Link' can be defined as a comparison pair that is accepted as being highly likely for the same individual, whereas a 'Non-Link' is a comparison pair that is not accepted as being highly likely for the same individual. [198]

However, in reality, it is not possible to determine exactly which comparison pairs are matches or non-matches; hence, it is usual to observe the combined number of comparison pairs. [198] Table 2.23 (page 119) adapted from Sayers et al. (2015), [199] and Blakely et al. (2002), [198] illustrates the classification of comparison pairs.

Table 2.23 Comparison of classification pairs, adapted from Sayers (2015) and Blakely (2002).

		True Match status	
		Match (same person)	Non-Match (different person)
Link status	Link	**a: True positive**	b: False positive
	Non-link	c: False negative	**d: True negative**

Additionally, the performance of RL in classifying the outcome is considered equivalent to a 'Match' in an epidemiological study and can be quantified in similar terms.[198, 199]

- M probability is equivalent to sensitivity = $\frac{a}{(a+c)}$
- U probability is equivalent to 1 minus specificity = $1 - \frac{d}{(b+d)}$

Sensitivity is the proportion of all records on one file that have a match in the other file that were correctly accepted as a link. Specificity is the proportion of all records on one file that have no match in the other file that were correctly not accepted as a link. [198]

The match pair ratio reflects the relative importance of a comparison field in predicting match status. [212] However, exact values for the parameters p (proportion of true matches), m, and u for each comparison field are unknown for a given dataset. [213] In addition, as datasets increase in size, false positives are more likely to occur. [214]

Step six and seven: To establish the match or non-match status of two records, the Fellegi & Sunter method produces a composite likelihood ratio. [209] The overall match weight is derived by calculating the ratio $log_2\left(\frac{m}{u}\right)$ for matches and $log_2\left(\frac{1-m}{1-u}\right)$ for non-matches. Additionally, this ratio is calculated for each identifier and then summed across all identifiers i.e.

$$Total\ Weight = (W_{surname} + W_{forename} + W_{date\ of\ birth} + W_{sex} + W_{postcode})$$

Record pairs with agreement on multiple identifiers will have large positive match weights; record pairs with disagreement on most identifiers will have negative match weights. [204]

Step eight: An example of the distribution of weights for matches and non-matches is illustrated in Figure 2.21 (page 121). Thus, PRL requires cut-off weights to be selected for classifying record pairs as links or non-links, which in turn determines the rates of missed-matches and false-matches. Typically, two thresholds are chosen, matched records with a composite weight above the assigned upper threshold are considered true matches; those below the lower threshold are considered unmatched (non-matches). Record pairs with weights falling between the thresholds are potential matches residing in the intermediate-weight 'grey area', and are subject to 'Clerical Review'. [202, 204, 211]

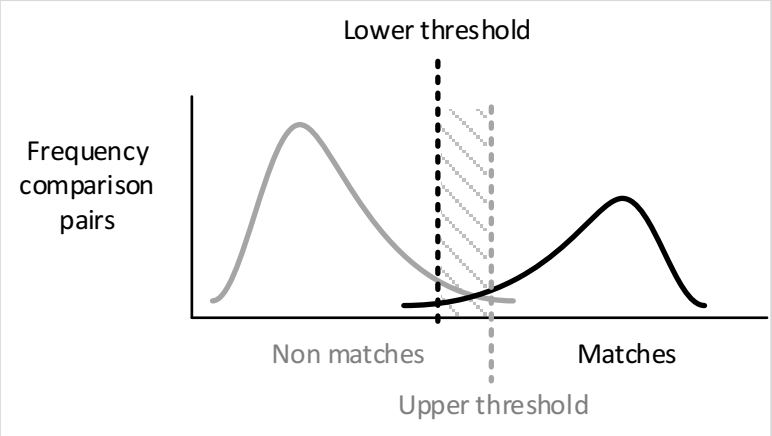


Figure 2.21 Thresholds used in probabilistic record linkage.

2.6.4 Complementary techniques to enhance RL

Record linkage is a key technological tool used to exploit the wealth of information contained in administrative and other data sources. However, it can quickly become difficult if there are errors in matching variables and/or when there is no unique, high-quality identifier is available, e.g. an ID number. [202]

Therefore, a review of the literature, identified 17 research articles, emphasising the use of complementary techniques to enhance DRL and/or PRL strategies, including studies that predominantly used quasi-identifiers e.g. a person's name, date of birth or address; as summarised within Table 2.24 (page 122).

Table 2.24 Research articles applying DRL and PRL techniques to demographic attributes.

Author(s)	Study setting	Data elements	Linkage	Software	Modification	Additional information
Blakely and Salmond, 2002 [198]	New Zealand Census & mortality datasets.	Geocode, sex, date of birth ethnicity, country of birth	PRL	SAS Automatch Software	None	Split date of birth (Day,Month,Year) Only one match required Blocking: Geocode
Boyd et al., 2015 [215]	Australian hospital admissions & mortality records.	Full name, sex, date of birth, address, dates for hospital events, mortality	PRL	Not stated	None	Blocking: Soundex surname, sex
Dean et al., 2001 [216]	USA Emergency medical services ED & inpatient hospital data	Patient sex, age, date of birth, hospital, admission date, ambulance run, dispatch codes	PRL DRL	SAS Automatch Software	None	Blocking: Sex, age, date of birth, admission date, hospital, ambulance run, and dispatch codes
Downing et al., 2005 [217]	West Midlands Ambulance Service & ED	Patient date of birth, sex, hospital arrival date/time	PRL	Not stated	None	DRL not possible as no unique identifier so used a combination of variables to create a unique identifier
DuVall et al., 2010 [218]	University of Utah Health Centre	Names (first, middle, last), date of birth, sex, social security number, telephone, address, city, zip code	PRL	Commercial software (no name)	Levenshtein edit distance	Expectation maximisation algorithm
Hermansen et al., 2009 [219]	US National Institutes of Health Diet & Health, Death Master File, Commercial list updates.	Both names, Date of birth Social security number, ID	PRL	SAS & WesMatch function	Jaro Winkler algorithm	Blocking: Soundex Surname
Kabudula et al., 2014 [220]	South Africa Health Facility.	Both names, date of birth, Village, Both names of another household member, ID	DRL PRL	Microsoft SQL Server	Jaro Winkler algorithm	Hybrid strategy: use DRL then PRL Expectation maximisation algorithm Combine: both names & date of birth Soundex: Forename, Surname Blocking: Sex, Year of birth, Village, Names
Karmel et al., 2010 [203]	Australia Pathways in Aged Care.	Both names, Date of birth, Sex, Postcode	DRL stepwise	Not stated	None	Concatenation of name, date of birth and sex to create own key as no common identifier Create a linked database

Author(s)	Study setting	Data elements	Linkage	Software	Modification	Additional information
Li B, et al., 2006 [221]	Canada Alberta Health Care Insurance, Vital Statistics, Hospital discharge summary	Both names, Date of birth, Sex	DRL	Not stated	None	Concatenation of name, date of birth and sex to create own key. Soundex: names if concatenated ID fails
Li X et al., 2014 [222]	Auvergne University France	Simulated data set Both names, date of birth, sex.	PRL	R package for Jaro Winkler	Jaro Winkler algorithm	Expectation maximisation algorithm
Méray et al., 2007 [223]	Dutch Perinatal Medical Registries	Mother & Child Date of birth, zip code, sex, birth weight, hour	DRL PRL	SAS	None	Comparative strategy: use DRL then PRL Blocking: Mother date of birth, zip code, child date of birth, sex
Morgan et al., 2016 [224]	England HES Data: Neonatal intensive care.	Baby: date of birth, sex, age, birth weight. Mothers delivery age, ethnicity, postcode, discharge date, previous pregnancies	PRL	Record linkage R-package	None	Expectation maximisation algorithm Manual review
Pacheco et al., 2008 [205]	Rio de Janeiro cohort database	Full name, date of birth, mother's name (maiden or married surname)	DRL	Python for windows	None	Soundex: Forename, Surname Hierarchical DRL algorithm
Poon et al., 2016 [225]	Teaching Hospital Hong Kong	Both names, age, sex, date of birth, ethnicity, district of residence, ID	PRL	Excel file	None	Sequence 1: PRL FS then JW Sequence 2: JW then PRL FS (not combined sequence)
Sayers et al., 2015 [199]	UK (hypothetical study)	Surname, forename, sex, education	DRL PRL	Stata, R	Jaro Winkler algorithm & Field relative frequency	Soundex: Forename, Surname Complex agreement patterns for names Posterior probabilities of a match, Blocking
Waien, 1997 [211]	Canadian Health Information and Ambulance records	Both names, age, sex, hospital, symptoms, severity of illness, nature of the emergency	PRL	SAS Automatch Software	None	New York State Intelligence Identification System (NYSIIS) Coding: Name Blocking: various combinations Creating longitudinal dataset
Zhu et al., 2009 [213]	USA Healthcare newborn screening.	Medical number, full name, gender, telephone, Zip code, city, date of birth, next of kin & doctor	PRL	Not stated	Field relative frequency	Weight modification enhances specificity with minimal decrease in sensitivity Split date of birth (Day, Month, Year) Blocking: Both names, Medical number

Furthermore, whilst Table 2.24 (page 122) provided an overview of each research article, i.e. study setting, data elements, type of linkage and software used; Table 2.25 (page 124) provides a high-level summary, identifying the most common linkage strategies and associated techniques employed by each respective author.

Table 2.25 An overview of common linkage strategies and techniques per author.

	Linkage strategy & techniques employed									
Author(s)	PRL strategy	DRL strategy	Blocking strategy	Soundex algorithm	Combination ID	Split Date of birth	String comparator	EM algorithm	Relative frequency	Count
Kabudula et al., 2014 [220]	✓	✓	✓	✓	✓		✓	✓		**7**
Sayers et al., 2015 [199]	✓	✓	✓	✓			✓		✓	**6**
Hermansen et al., 2009 [219]	✓		✓	✓			✓			**4**
Zhu et al., 2009 [213]	✓		✓			✓			✓	**4**
Blakely and Salmond, 2002 [198]	✓		✓			✓				**3**
Boyd et al., 2015 [215]	✓		✓	✓						**3**
Dean et al., 2001 [216]	✓	✓	✓							**3**
DuVall et al., 2010 [218]	✓						✓	✓		**3**
Li B, et al., 2006 [221]		✓		✓	✓					**3**
Li X et al., 2014 [222]	✓						✓	✓		**3**
Méray et al., 2007 [223]	✓	✓	✓							**3**
Downing et al., 2005 [217]	✓				✓					**2**
Karmel et al., 2010 [203]		✓			✓					**2**
Morgan et al., 2016 [224]	✓							✓		**2**
Pacheco et al., 2008 [205]		✓		✓						**2**
Poon et al., 2016 [225]	✓						✓			**2**
Waien, 1997 [211]	✓		✓							**2**
Count	**14**	**7**	**9**	**6**	**4**	**2**	**6**	**4**	**2**	

2.6.5 Linkage strategy

Within the majority of studies, there was no common identification number available for linkage. Hence, the data linkage strategies employed by most authors included the application of either DRL or PRL strategies, whilst one study [216] compared the result of applying each strategy separately, other studies [220, 223] utilised a hybrid approach of applying DRL followed by PRL strategies, as this improved sensitivity and yielded optimal results.

Furthermore, study [220] used a two-step approach; step one linked the datasets using DRL, followed by the application of step two using PRL on remaining records. However, the authors should have considered taking a further step, undertaking PRL on all the records as a comparator, as this approach would identify any missed matches from step one.

Likewise, study [223] reported using exactly the same eight data elements to compare both DRL and PRL strategies, in order to elucidate which technique was prolific. The authors reported using the DRL strategy, i.e. exactly matching on all eight data elements, resulted in the lowest number of record links (n=41,673). In contrast, employing PRL resulted in an 80% increase in the number of record links (n=76,050). Thus, the authors reported that PRL outperformed DRL strategies.

Nonetheless, the specific linkage strategy employed will be determined ultimately by various factors, including the availability of data elements, amount of data to link, provenance, levels of DQ, technical capabilities of applications, and influenced by the level of knowledge/understanding of RL strategies possessed by the researcher.

2.6.6 Blocking strategy

PRL results in the Cartesian product (described in Table 2.22, page 117) of all possible record pair comparisons. Therefore, the application of blocking strategies in which blocking fields agreed, enabled the division/partitioning of the Cartesian product into smaller more manageable subsets [213] and was the most commonly applied supporting technique to RL. Thus, the application of blocking techniques help reduce the number of comparison pairs and increase the efficiency of RL strategies. [198]

The majority of authors report that ideal blocking data elements should contain a high variety/range of values and very low missing values, for example, the data element ‘Surname’ will contain variety

(from A to Z), and be more like to be recorded so the DQ dimension of completeness will be relatively high. Hence, blocking strategies increase the proportion of true matches among possible pairs while decreasing the number of comparisons, making the RL process more efficient and manageable. [210]

2.6.7 Soundex algorithm

The Soundex algorithm is a 'name phonetic algorithm' and encodes names using alphanumeric codes. [199] The algorithm associates numbers with different groups of consonants, producing an initial letter followed by a numeric code, which adapts to variations in names that sound alike. [221]

Therefore, when considering a person's surname, the Soundex algorithm is coded based on the way the surname sounds rather than how it is spelt, e.g. SMITH and SMYTH, although these surnames are spelt differently they sound the same and are therefore allocated the Soundex code 'S530'. [226] In contrast, the word 'swan' has the code S500, and 'umbrella' has the code U516, both codes are different. Thus, the use of approximate comparators is an effective way to improve RL results when field values contain misspellings, alternate spellings, and typographical errors. [218]

Most authors applied the Soundex algorithm to obtain corresponding codes for either surname, forename or both names. For example, one author [205] used the Soundex algorithm to obtain the phonetic code for all person names within their dataset, which was used as a blocking strategy by grouping records into comparison blocks, i.e. Dataset-A-S530 linked to Dataset-B-S530. This strategy helped to prevent the exponential growth of processing time for RL.

Furthermore, one study [221] reported using the Soundex algorithm to generate a Code to be used as an additional comparator between names that fail to match exactly due to variations in spelling, e.g. comparing the names Smith and Smyth. Such a comparison would result in no match. However, as both names possess the same Soundex code (S530) the result is a match.

Therefore, Soundex codes can provide additional information when comparing name data elements and provide an effective strategy for blocking purposes. However, a caveat when undertaking PRL on names is to not 'double-count' by first, accepting the match based on names and second, the match of Soundex codes based on the same names, as this introduces an element of bias. [202] Hence, blocking data elements should be selected and applied in a precise manner.

2.6.8 Combination ID

The main objective of RL strategies is to facilitate the identification of the same person across two or more datasets, which is reasonably straightforward if there is a unique identifier available, enabling the application of DRL. Conversely, if there is no such unique identifier common to both datasets, the application of DRL techniques is not possible, and RL becomes more challenging.

One potential strategy employed to address such instances is to consider using a combination of quasi-identifiers to enable the unique identification of an individual. [197] Within this context, 'quasi-identifiers' refer to any additional demographic information available within the dataset, i.e. forename, surname, date of birth or address. [202] Hence, multiple quasi-identifiers can be combined to create a new 'Combination ID', which facilitates the linkage of datasets via DRL strategies, e.g. Dataset-A-CombinationID ≡ Dataset-B-CombinationID.

The majority of studies incorporated additional person identifiable data elements, e.g. the authors [221] used four quasi-identifiers (surname, forename, sex, and date of birth) within their study and stressed the importance of formatting the identifiers exactly the same way across all datasets, i.e. capitalising letters, removing blank spaces and symbols.

Additionally, the authors [221] reported using a combination of three quasi-identifiers (surname, sex and date of birth) to create a Combination ID was optimal to achieving the best linkage results. However, upon reflection, the approach used by the authors [221] will inevitably lead to a higher number of false matches, especially in linkages where the surname is commonly used, e.g. Smith. Therefore, another approach would be to consider including the forename within the Combination ID as this would improve the correct identification of a person and linkage of records.

Likewise, one study [220] used a combination of quasi-identifiers (forename, surname and date of birth), whereas, in another study, the authors [203] created a new ID based on the concatenation of selected letters from person name, date of birth and sex.

Moreover, personal characteristics fixed at birth very rarely change, e.g. sex, ethnicity, date of birth, place of birth. Proper names rarely change during the lifetime of a person, e.g. birth surname, first forename, although surname changes usually occur through marriage. Social demographics, e.g. address and postcodes can change more frequently. [202]

Conversely, the authors [217] created a unique identifier from a combination of hospital name, date of birth, sex, and arrival date/time. This demonstrates that not all quasi-identifiers have to be based on personal demographics.

Nevertheless, in the absence of a unique ID, the authors have revealed that using carefully selected quasi-identifiers can facilitate RL and the identification of the same person across the numerous datasets. The key finding was that using more quasi-identifiers improved the chances of correct matches, i.e. the success of RL relied on choosing an appropriate combination of common identifiers with a good level of DQ (completeness). [221]

2.6.9 Split Date of birth

Two studies [213] and [198] applied a technique to split/disaggregate the date of birth into its component parts, i.e. Day, Month and Year. The advantage of this approach is that first, it yields more data elements to use in matching records and second, it allows for any transcribing errors.

Consequently, transcribing errors can occur, especially in situations where the quick recording of information is required. For example, within a Hospital ED setting during busy periods, reception staff are under pressure to register patients quickly, so a date error such as typing 14/02/2017 instead of 15/02/2017 may occur inadvertently. Therefore, by splitting the date of birth, at least two out of the three components, Birth Month and Birth Year will result in a correct match.

2.6.10 Relative Frequency

In 1959, Newcombe [201] introduced methods for using the specific values or relative frequencies of occurrence of fields such as surname. However, as the Fellegi-Sunter [209] approach does not explicitly accommodate the notion of a value-specific weight, each data element receives an identical agreement weight for all record pairs.

For example, the intuitive idea is that if surnames such as ‘Vijayan’ occur less often than surnames such as ‘Smith’, then ‘Vijayan’ should have more distinguishing power. [227] Thus, uncommon surnames will adjust the weight upward, whilst common surname should adjust the weight downward for that record pair. [228]

Correspondingly, the authors [213] and [199] support the notion that the relative frequency for either rare or common data items, should be incorporated into the matching process, because agreement on a common value should not be scored the same as agreement on an uncommon value for that field. The author [213] further reports that frequency-based modification served to enhance the performance of the Fellegi & Sunter algorithm, by eliminating false-positive matches and improving specificity.

2.6.11 String comparator

When matching records, typographical errors will affect performance when comparing two strings exactly (character-by-character) [197], e.g. Smith and Smyth. The author Jaro [229] devised a string comparator function, which represents degrees of partial agreement between the values zero and one, and accounts for insertions, deletions, and transpositions between two strings.

The Jaro algorithm (Figure 2.22 page 129) has three components, first, compute string lengths, second, find the number of common characters in the two strings, and third, find the number of transpositions. The definition of 'common' in this context is that the agreeing character must be within half the length of the shorter string. [197] The function ϕ (phi) represents an exact agreement when it takes value one and represents partial agreement when it takes values of less than one. [229]

Jaro algorithm	$String\ s_1 (with\ length\ len_{s1})$ $String\ s_2 (with\ length\ len_{s2})$ $N_C\ is\ common\ characters\ between\ s_1\ and\ s_2$ $N_t\ is\ the\ number\ of\ transpositions$
$\Phi_j(s_1, s_2) = \frac{1}{3}\left(\frac{N_C}{len_{s_1}} + \frac{N_C}{len_{s_2}} + \frac{0.5N_t}{N_C}\right)$	

Figure 2.22 overview of the Jaro algorithm.

Winkler [230] identified that fewer errors were made at the beginning of the string than at the end of the string. [202] A separate study by authors Pollock et at.,(1984) [231] provided empirical evidence that quantified how the probability of keypunch errors increased as the character position in a string moved from the left to the right.

Additionally, Winkler (1990), [230] demonstrated how a variant of the Jaro string comparator phi ϕ dramatically improved matching efficacy, compared to situations when string comparators are not used, and therefore enhanced the Jaro string comparator by introducing weights to 'up-weight'

similarity at the beginning of the string. [199, 202] The Winkler variant, referred to as the 'Jaro–Winkler' string comparator, is widely used in computer science. [197]

Thus, Jaro-Winkler methods [230] compare the number of common characters and character transpositions between two strings. For example, comparing strings using the Jaro-Winkler algorithm:

- 'Compatible' with 'Compatibility' results in a score of 0.91
- 'Swan' with 'Sparrow' results in a score of 0.64
- 'Swan' with 'Umbrella' results in a score of 0
- 'Swan' with 'Swan' equals a score of one, indicating an exact match

Moreover, the authors [220] reported using Jaro-Winkler string comparison techniques, whereas, authors [225] compared the independent application of Jaro-Winkler and Fellegi-Sunter algorithms. The authors [199] report that in scenarios where the relative frequency of data elements are being used, e.g. name comparisons, then the Jaro-Winkler algorithm readily facilitates incorporation and match weight calculation.

2.6.12 Expectation Maximisation algorithm

The Expectation Maximisation (EM) is a method to calculate match probabilities using maximum likelihood estimates of parameters in probabilistic models with incomplete and/or missing data. [222] Using a subset of data, it seeks to estimate the proportion of true matches to establish the match to non-match threshold. [213]

The EM process has two steps, first the Expectation step whereby the model is run using current parameter estimates, and second, the Maximisation step in which parameters are re-estimated to maximise the likelihood of the observed output of the model. Thereby estimating optimal parameters for a model over several iterations, until the parameters converge. [218]

However, in practice, the true number of correct matches is unknown, [222] and there is no assurance the EM algorithm will divide the product of $A \times B$ into classes that correspond to M (match) and U (non-match) probabilities.

Thus, caution must be observed when applying the EM algorithm to real-world data, as it does not always yield reasonable parameters due to the moderately high proportion of truly matching pairs that disagree on names or addresses. [197]

2.6.13 RL Software selection

The studies revealed that each author had chosen a specific software package to utilise in order to undertake RL using DRL and/or PRL strategies. Whilst some authors decided not to disclose what commercial software they used, others reported using the following software packages:

- **Proprietary software**: SAS [232]; Microsoft SQL Server [233]; Microsoft Excel [234]; Stata [235]
- **Open-source software**: R for Statistical Computing [236]; Python Programming Language [237]

Additionally, each of the software packages utilised for the respective studies enabled the author(s) to create bespoke algorithms, so that authors(s) maintained complete control of the RL process. Conversely, some commercial packages adopt a 'Black-box' approach [238] whereby the underlying RL algorithms and allocation of threshold values are hidden, and users remain unaware of which parameters were applied.

Therefore, when selecting an appropriate software package for potential studies, researchers must carefully consider various factors, e.g. technologies available; software cost/benefits; operating system compatibility; specialist training requirements; data quality/size and software adaptability to create bespoke algorithms. Ultimately, choosing an appropriate software for RL purposes remains the decision of respective author(s) and to what degree they consider the selected software package fit for purpose.

2.6.14 Application of RL strategies

Whole Systems Integrated Care (WSIC): is an integrated data source encompassing a population of 2.4 million patients within North West London, and contains data from over 400 provider organisations that include 360 GP Practices, two Mental Health Trusts, two Community Trusts and eight acute providers. [239, 240] DRL strategies are employed, utilising the NHS Number as the unique identifier to enable the linkage of numerous patient health and social care activity datasets. [240] Next, a unique Master Patient Index (MPI) is created, and all integrated records are stored

within a secure data warehouse [239] linked to interactive Dashboards and made available to healthcare professionals for further analysis/visualisations. [239]

Clinical Practice Research Datalink (CPRD): has been providing anonymised longitudinal primary care records for public health research since 1987, to improve drug safety, best practice and clinical guidelines. A key feature of CPRD is that it facilitates anonymised primary care individual-level patient data to be linked to secondary care and other health and area-based datasets, e.g. all HES datasets, Death and Cancer registration, Mental Health and Index of Multiple Deprivation. [241, 242] The CPRD datasets are linked to other datasets using the same step-based DRL strategy employed by HES, [206, 243] and detailed in Table 2.21 (page 116).

The Oxford Record Linkage Study (ORLS): established in 1963, employed a hybrid approach, by applying DRL strategies using the NHS number and adapted Newcombe's [201] PRL methodologies using quasi-identifiers (surname, forename, gender, date of birth, address, postcode) to link all hospitalisations, births and deaths for the 2.5 million resident population of the County Borough of Oxford. [242, 244, 245]

Western Australian Health Services Research Linked Database: used the records of approximately 1.7 million people and linked multiple health and administrative datasets to assign geocodes at the population level that support health planning, evaluation and research. Linkage achieved via the use of person and quasi-identifiers and PRL strategies. [242, 246]

Secure Anonymised Information Linkage Databank (SAIL): used in Wales, and holds over 500 million records including hospital and general practice data. The authors [247] used a SQL-based algorithm to apply DRL and PRL methods to a set of matching variables which include the NHS number and five quasi-identifiers first name, surname, postcode, date of birth and gender. The DRL algorithm looked for an exact match on all five variables, whilst the PRL approach employed of a variety of techniques, e.g. Lexicon matching and Soundex matching to identify matching records, thereby establishing an infrastructure for future research studies. [242, 247, 248]

2.6.15 Section summary

This section provided an overview of strategies that facilitated the RL process, by describing and providing examples of the two main approaches, i.e. DRL (exact matching) and PRL (formalised by Fellegi & Sunter). In addition, research articles that placed emphasis on the use of complementary

techniques to enhance DRL and/or PRL strategies, and studies using quasi-identifiers in the absence of a unique identifier common to both datasets were summarised. Furthermore, specific corresponding strategies were thematically identified and discussed, e.g. Blocking, Soundex, String comparators. Thereafter, an overview of software appropriate for RL was provided, and a brief description of the HES database and the associated step-based DRL approach it employed was presented, and finally, examples of the application of RL were outlined.

2.7 Chapter summary

Overall, this literature review sought to provide the reader with appropriate background about the study setting, and identify relevant concepts, theories, methodologies and research within the field of Medical Informatics, to address the issue of health information fragmentation due to disparate CIS and consequent data silos.

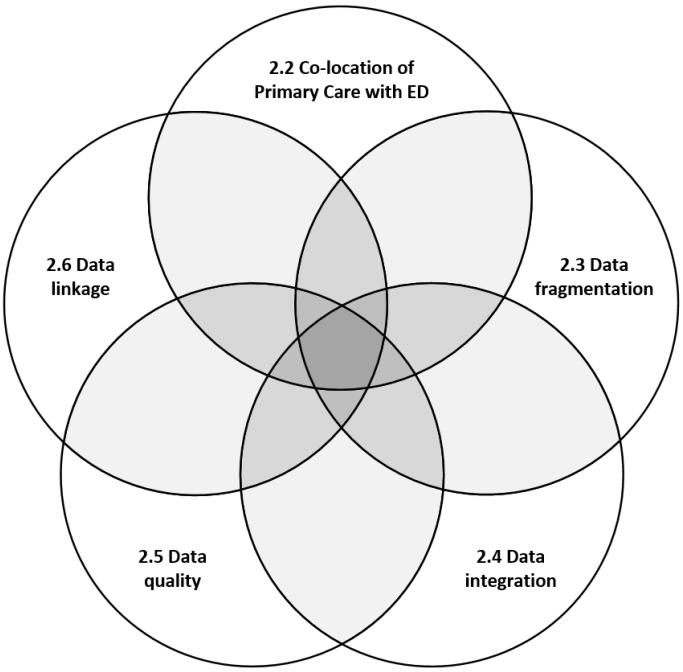
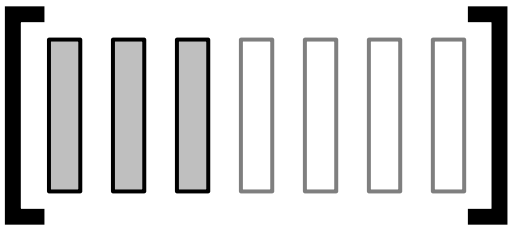


Figure 2.23 Interconnection of chapter sub-sections

Moreover, the findings from each section were used to help identify the gap in knowledge, and guide this research in specifying the research questions, aims and objectives for this thesis, as presented in Chapter 1, page 16.

Chapter 3 Research Framework & Methodology



3.1 Introduction

This chapter presents the theoretical framework underpinning this study and contextualises the issue of information fragmentation within this theory. Furthermore, it presents the conceptual framework developed from informatics techniques that provides an end-to-end approach for data linkage and integration. The research methodology and philosophical underpinning for this study are presented along with appropriate background about the research setting.

3.2 Theoretical framework

A Theory is a set of interrelated concepts/statements that provide a systematic view of a phenomenon. [249, 250] Theories enable the proposition of relationships, formulation of predictions, and influence strategies for action. [249] Concepts are terms that abstractly describe the phenomenon by providing meaning or identity to facilitate the presentation of ideas relevant to theory. [250]

Thus, a theory enables the selection, classification/organisation of concepts to explain an issue by formulating/identifying causal relationships that seek to understand how/why observed phenomena occur, in order to predict outcomes. Correspondingly, theoretical frameworks connect researchers to existing knowledge in the field; help to clarify important facts, evidence, questions; provide a frame of reference; and a basis for hypothesis formulation and testing. [251]

3.2.1 Systems Theory

A review of the literature identified 'Systems Theory' as the overarching theory underpinning this research. The importance of 'systems' was recognised by Ludwig von Bertalanffy (1901-1972), who during the 1930s and 1940s, developed the concept of General Systems Theory (GST), which represented a 'complex' of interacting elements or component parts that together created a functioning whole or entity. [252, 253]

Systems may be living or non-living and can include computer systems, school systems, the healthcare system, and a person/human. [254] For example, 'Humans' are comprised of biological, physiological and psychological elements, whereas 'Families' are comprised of different members as elements (i.e. mother, father, sister, brother, daughter, son) with different types of relationships as elements (i.e. spousal/couple, parent-child, sibling). 'Open' systems are receptive to and influenced by outside stimulus, whereas 'Closed' systems are solid/unreceptive to outside stimulus. [252]

Within a system, the different elements do not function in isolation; rather, they function by depending on and interacting with each other to complete the system as a whole. Therefore, the 'whole' cannot be complete without the presence/participation of each of the elements. [252]

GST seeks to explain the complex interplay of various parts of a system, and how these different parts interrelate, [252] by examining the interacting components, reciprocal relationships and mutually inter-influencing factors that affect behaviour holistically. [253]

For example, the aim of GST is to observe a person in their environment, conduct a holistic assessment of the individual, family, group and/or community, and explore all possible areas that may require or benefit from intervention. [252, 253] The basic assumptions of GST include [252]:

- The whole system is greater than the sum of its individual parts. Each system consists of several interlocking elements/subsystems that interact together to form a functional purposeful whole, within which each element/subsystem has a purpose.
- The parts of a system are interconnected and interdependent. Any change/movement in one part of the system will result in a change/movement in other parts of the system, Figure 3.1 (page 136) adapted from Teater (2014), [252].

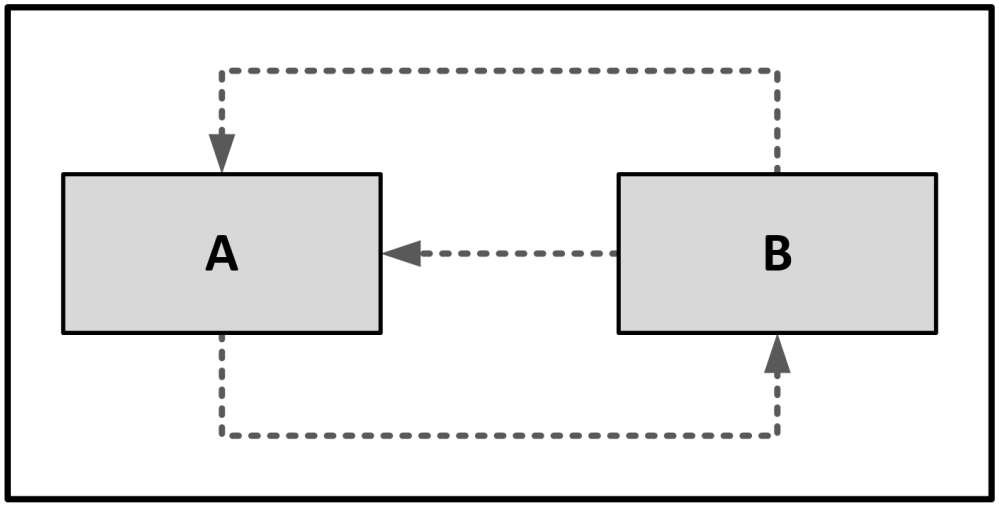


Figure 3.1 System interconnection and interdependence (adapted from Teater (2014)).

- A system is either directly/indirectly affected by other systems. Within a system, different elements/subsystems interact/affect each other, in a similar manner, systems interact/affect other systems.
- All systems have boundaries. The system boundary distinguishes it from other systems; however, in some scenarios, the boundaries may overlap with other systems.

- All systems need to remain in a state of equilibrium/homeostasis. Despite conflicting influences, the goal is to maintain internal balance.

However, GST has some limitations, for example, situations in which it is not possible to conduct accurate assessments of all the systems/variables will result in a lack of information about how or to penetrate a system to address issues. Resources and time factors, i.e. in situations where an assessment reveals several systems that require intervention, however, time/resource constraints permit tackling only the most prominent issue. In addition, not all systems recognise the need, nor are they open to any form of intervention. [252]

Nonetheless, GST maintains that the whole of a system is greater than the sum of its parts, and emphasises the need to assess the interconnection and interdependence of systems, to identify how these underlying interactions contribute to problems. Thus, GST seeks to restore equilibrium by intervening in one/more systems to mitigate the presenting problem. [252]

3.2.2 GST & Information system performance

When applying GST to Information Systems, the authors Whitworth et al., 2008 [255] state: "A system is an entity within a 'world', whose nature defines the nature of the system. The world need not be physical, so if information, cognitive, and social 'worlds' exist, then information systems, cognitive systems, and social systems, respectively, can exist within them."

Modern information systems face performance demands similar to biological systems. [256] System performance, involves multiple aspects, linked by the elements of a boundary (extendibility, security), the internal structure (flexibility, reliability), effectors (functionality, usability) and receptors (connectivity, privacy). For example, the Web of System Performance (WOSP) model assumes a GST perspective, so the term 'system' can be applied at both technical and social levels. [255, 256]

The WOSP model imagines system properties as connected by elastic so that increasing any one property puts tension on its opposite and to a lesser extent all the other properties. For example, referring to Figure 3.2 (page 138) adapted from Whitworth et al., 2008 [255], which illustrates the goals Flexibility and Reliability opposite one another. However, as the endpoints on the web are merely in tension, opposing goals pulled together in opposite directions will expand the web. [256]

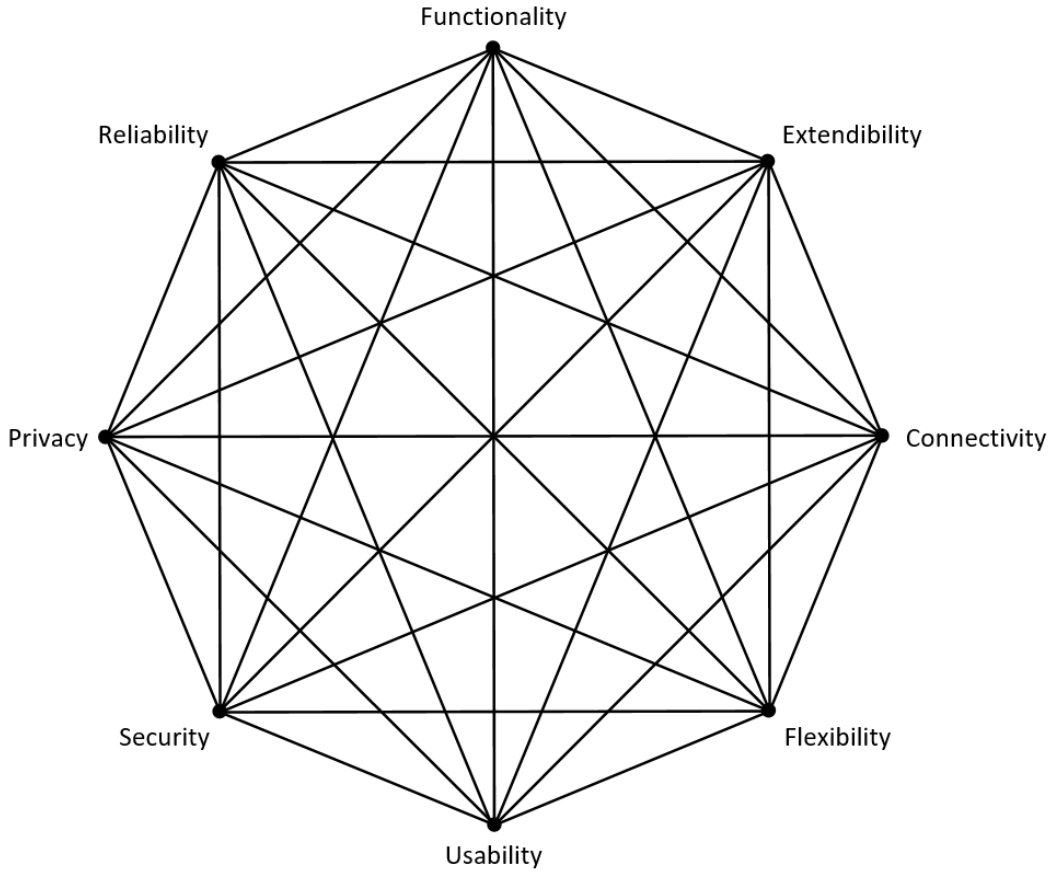


Figure 3.2 The Web of System Performance (adapted from Whitworth et. at (2008)

Additionally, the web area represents overall system performance, the larger the area, the more the performance potential. The shape of the web shows the contribution of each goal to overall performance in a given environment, e.g., turbulent environments tend to favour reliability. The lines within the web show goal tensions imagined as rubber connecting bands, e.g., improving flexibility may unintentionally reduce reliability. [255]

Thus, when designing systems to satisfy multiple purposes, nature provides an excellent guide and humans, a good example. One must take a realistic view of disparate system goals in 'design tension' within a changing environment, as the computer systems we create must also work with the systems that we do. [256]

3.2.3 Sociotechnical Systems Theory

Over the past few decades, there have been significant increases in the amount of Healthcare Information Technology (HIT) used to support various elements of work within healthcare systems. Along with the increased efficiency and power of technology, there has been a corresponding recognition of the inherent complexity of healthcare environments in which humans and technology interact. [32]

Historically, health technology was evaluated in conceptual isolation from the context in which it operated. However, newer perspectives within the health sciences domain support conceptualising healthcare environments as dynamic systems that contain both social and technical, i.e. sociotechnical entities, commonly known as Sociotechnical Systems (STS). [257]

The term 'Sociotechnical' was introduced in the 1950s by the Tavistock Institute, whereby a Sociotechnical System (STS) is a social system sitting upon a technical base and should be considered as the whole unit, not social and technical systems side-by-side. The use of 'email' is an example of social communication by means of technology. [258]

Additionally, the words 'Socio' (of people and society) and 'Technical' (of machines and technology) are combined to give the term 'sociotechnical', which refers to the interrelatedness of 'social' and 'technical', [259] and describe the complex issues related to the interconnected nature of society and healthcare technology. [260]

The STS perspective recognises organisations as 'work systems' with two autonomous but interconnected parts which need to be considered jointly, first the social system which concerns the relationships among people, and second the technical which concerns the processes, tasks, and technology required to transform inputs into products/outputs. Since social and technical systems are interdependent/intertwined, it is important to consider interactions between them and the context in which they are embedded/used. [261]

Thus, the STS framework considers technical, human, social, and organisational factors in the design of organisational systems, [261] and STS research focuses on how social and technical aspects integrate into a higher-level system with emergent properties. [258]

The STS body of knowledge aligns with viewing reality from a balanced perspective that appreciates the importance and agency of both humans and technology, and all things, e.g., people, technology, processes, have the potential to interact with each other to mutually inform, influence, and generate action. [257]

Furthermore, adopting an STS approach has implications for the development and evaluation of HIT applications in work practices. For example, it may influence thoughts/views regarding the concept of medical information, or the way in which HIT is embedded within organisations, whilst taking a critical approach of 'recipes' for successful system development. [257]

Healthcare practices are heterogeneous networks from an STS perspective, i.e. work practices conceptualised as networks of people, tools, organisational routines, documents. [257] The process of Emergency Medicine is a 'sociotechnical' process, whereby each action/activity is governed by information that facilitates the coordination of ED professionals with numerous resources and technologies. [262]

The core activity of any ED healthcare practice is the effective management of patients, e.g. investigations, monitoring, intervening and re-intervening in order to treat patients. Additionally, organisational complexity and the unpredictable nature of patients' reactions to interventions, result in an ongoing stream of sudden events that need dealing with on the spot, by whoever happens to be present, with whatever resources happen to be at hand. [257] Thus, healthcare work entails ongoing negotiations about the nature of tasks and relationships between those who execute the tasks. [257]

Nevertheless, the behaviour of any system that transpires from the interaction of its components will be affected by increasing the number of components, making it more difficult to predict the outcome of even seemingly simple changes. Within complex systems such as healthcare, simple changes/fixes will always result in unexpected consequences. Ultimately, the web of interactions that are required to make anything work in a complex organisation inevitably involve humans solving problems with limited resources and working around imperfect processes. [263]

Moreover, EDs contain an interrelated assembly of humans and 'things', that work synergistically to deliver patient care. [257] The flow of information from an 'activity' determines subsequent

workflow, [262] and any interruption in information flow will inadvertently affect clinical decision-making, patient care, and departmental performance.

Coiera (2004) [263] states from an STS viewpoint, technical systems have social consequences, i.e. the introduction of new technology into a setting will affect its intended users whilst indirectly impacting on other people/services. Healthcare systems are inherently complex, and to understand STS, a greater understanding of the interactions between people and technology is required.

3.2.4 Interactive Sociotechnical Analysis framework

A sociotechnical lens on everyday action can provide researchers with a dynamic perspective to view processes and examine the role of technological systems within healthcare. [257, 263, 264] The following example serves to illustrate this statement by focusing on the interconnection and interdependence of systems.

The interactions between HIT and a healthcare organisations STS, e.g. workflows, culture, social interactions, and technologies result in many unintended/undesired effects, and often interplays between new technologies and existing sociotechnical conditions are overlooked. [31]

Thus, the authors Harrison et al. (2007) [31] frame their study of unintended consequences of HIT, in a sociotechnical lens via the use of an Interactive Sociotechnical Analysis (ISTA) model, which provides a framework identifying important relationships amongst new HIT, workflows, clinicians, and organisations. [32]

From a sociotechnical perspective, all actors within the situation are interconnected and interdependent on each other to generate/perpetuate action. [32] The ISTA framework emphasises the interaction/interdependence among social and technical systems and recursive relations among sociotechnical subcomponents, which underlies the potential for unintended consequences. [31] Figure 3.3 (page 142) adapted from [31] illustrates the interactions among the main subcomponents.

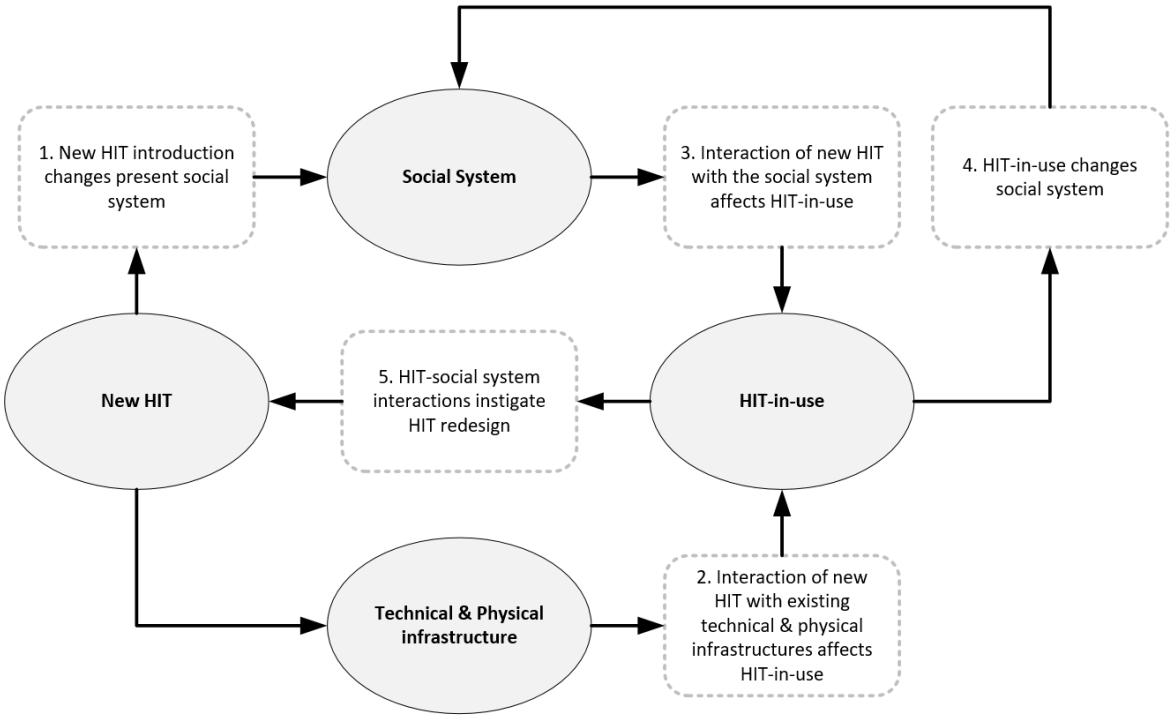


Figure 3.3 Interactive sociotechnical systems, adapted from Harrison et al. (2007).

Further detail describing the types of interactions within the ISTA framework, adapted from Harrison et al. (2007), [31] and Booth (2015), [32] is provided within Table 3.1 (page 143).

Additionally, the authors Sittig et al. (2010), [264] affirm that Harrison et al.'s [31] ISTA framework provides an excellent broad overview of the complex, emergent, interrelationships between the HIT, clinicians, and workflows within any healthcare system.

Nevertheless, as new HIT intertwines with the existing sociotechnical system, many important consequences of sociotechnical interactions emerge that cannot be fully anticipated at the design/implementation stages. [31]

For instance, when implementing a new service and CIS, there are first, many potentially important social elements, e.g., workflow, staffing schedules, clinician types, policies. Second, technical elements, e.g., number of pre-existing systems, degree of interoperability, interface, usability, all of which are required to come together in a synergistic fashion.

However, if some elements, whether social or technical or both, should fail or be resistant, other elements of the implementation may also suffer. [32]

Table 3.1 Types of interaction within the ISTA framework (adapted from Harrison (2007) and Booth (2015))

Interaction	Detail
Type 1	Introduction of new HIT changes the present social system. • Generating new/more work for clinicians. • Changes to communication patterns. • Other workflow changes/disruptions.
Type 2	Interaction of new HIT with existing technical and physical infrastructures in the environment affects HIT-in-use. For example: • The physical environment may not be suitable for HIT use. • Other HIT already present may not be compatible with introduced HIT. • Technical/physical infrastructure issues mediate how a new HIT is actually used. • HIT may deviate from its original design when implemented into clinical areas. • Pre-existing technical/physical infrastructures may influence HIT.
Type 3	Pre-existing social systems facilitate the reinterpretation of HIT in-use. • Clinicians may prefer using paper-based records. • The pre-existing social system favours the use of paper records. • Implementation of EMR to eliminate the use of paper may result in workarounds to keep the paper actor present within clinical work processes.
Type 4	HIT-in-use can influence and change pre-established social systems. • Implementation of HIT results in changes of clinician's roles via the addition or removal of tasks, compared to their previous roles. • HIT can evolve pre-existing social systems by changing power structures in the environment, e.g. rewriting of policy and best practices, • Clinicians may become overly dependent on technology in certain automated situations, e.g. use of clinical decision support tools.
Type 5	HIT-in-use and social system interactions can engender overall HIT redesign. • As all clinical practice occurs within a sociotechnical environment, it is clear that the addition of a new HIT will eventually influence both the social and technical features of the environment. • HIT-in-use and social systems eventually force an evolution of each other in potentially asymmetric, yet reciprocal fashions, with unpredictable outcomes.

Although technologies such as CIS possess great potential for benefitting medical care, they must interact with the work system as a whole, including people, work practices and other technologies, to enable or hinder the coordination of dynamic work demands. [265]

Therefore, the optimal utilisation of any CIS is extensively dependent upon the precise interrelation and alignment with the system, and healthcare professionals. [257] However, the STS approach does not provide a 'solution' to the many problems of CIS development; rather, it provides a viewpoint on healthcare practices focusing on empirical, qualitative insights that serve to improve analytical understanding. [257]

The goal of achieving interoperability is one of the many STS challenges faced by healthcare organisations. Within the context of STS, interactions implied by interoperability include, first, technical interactions, e.g. data exchange, inter-networking. Second, nontechnical interactions relating to inter-organisational initiatives, e.g. service coordination, cooperation, collaboration or alignment. The latter are factors concerned with service 'co-operability' and can act as barriers to holistic interoperable solutions. [266]

Thus, to achieve interoperability within STS environments, the mitigation of nontechnical barriers to interoperability becomes prerequisite to implementing technical interoperability solutions. [266] However, the extent of STS issues become particularly apparent during change processes, e.g. when implementing a technical system, the technological context can influence/impact social relationships, and vice versa. [260]

Nonetheless, as social systems have technical consequences, similarly, technical systems have social consequences. [263] Hence, adopting an STS perspective offers a powerful lens to evaluate complex healthcare systems and HIT interaction, by seeking to understand the nature of the relationships and level of interaction between these interdependent systems. [252, 267]

3.2.5 The relevance of STS to this research

One of the key assumptions of General Systems Theory is that the whole system is greater than the sum of its individual parts, each of which is interconnected and interdependent. [252] A system is an entity within a world (which need not be physical); thus, information systems exist within the world of 'information'. [255]

Healthcare environments can be conceptualised as dynamic systems that contain both social and technical entities, regarded as Sociotechnical Systems. [257] Emergency Medicine is a sociotechnical process [262] and healthcare practices considered as heterogeneous networks, [257] in which each action/activity is governed by information to enable coordination of EM professionals with resources and technologies. [262]

This research, underpinned by STS theory, investigates how the introduction of new HIT results in changes within the existing social system, affecting technical communication patterns. [31] As the flow of information from an 'activity' determines subsequent workflow, [262] any interruption in

information flow will inadvertently affect clinical decision-making, patient care, and departmental performance.

Hence, the introduction of new HIT results in changes to the existing social system, and the interaction of new HIT with existing technical/physical infrastructures in the environment affects HIT-in-use. [31, 32] However, it is not possible to anticipate all the consequences of sociotechnical interactions that emerge at the implementation stage. [31]

For example, taking an STS perspective, all actors are assumed to be interconnected and interdependent, [32] and within the technical context, 'actors' refers to HIT and the CISs within the system. Therefore, technical interactions between actors, i.e. data exchange and inter-networking will affect the degree of interoperability between pre-existing and new HIT. [32, 257] Consequently, achieving interoperability is just one of the many STS challenges faced by healthcare organisations. [257]

Nonetheless, a lack of technical interoperability inevitably results in the fragmentation of healthcare information leading to the formation of data silos, an issue that this research seeks to investigate. Using an STS lens, Figure 3.3 (page 142) [31] illustrates potential areas for unintended consequences of HIT amongst subcomponents.

Although the STS approach does not provide a solution, it does provide a viewpoint on healthcare practices focusing on empirical, qualitative insights that serve to improve analytical understanding. [257] Therefore, recognising STS within EM and the respective interactions between system elements is crucial to gaining an understanding of organisational and technical factors that contribute to information fragmentation. [268]

3.3 Conceptual framework

Theories are abstract, generalisations that explain relationships between concepts and phenomena, and theoretical frameworks are the general ideas that underpin conceptual relationships, but are not specific to individual studies. [251] However, theories can be captured in the form of a conceptual framework. [269]

The authors Miles et al. (2014), [270] state a conceptual framework "explains, either graphically or in narrative form, the main things to be studied, the key factors, variables, or constructs and the presumed interrelationships among them".

Conceptual frameworks are specific to the piece of research being undertaken and provide an outline of the sequence, relationships and direction of concepts [251] for the phenomenon being investigated. [271] Additionally, they enable the prediction and explanation of outcomes/behaviours [269] and provide a justification for the research. [271]

Thus, a conceptual framework is a set of "interrelated concepts that represent an image of a phenomenon" [249] and can be presented as a diagram that depicts the "wholeness" of a phenomenon. [250]

3.3.1 The conceptual framework for this research

The phenomenon of interest for this research is the fragmentation of health information caused by data held in silos within multiple disparate CIS. Hence, this research was viewed via a sociotechnical lens, whereby the implementation of a new service and accompanying new HIT affects existing sociotechnical systems within an organisation, in which all actors are interconnected and interdependent on each other. [32]

However, therein lie the potential for unintended consequences of HIT. [31] From a technical perspective, pre-existing ICT infrastructures do not support nor interact with new HIT, i.e. they are not interoperable, resulting in the formation of data silos. Thus, leading to the discovery of the invisible problem of information fragmentation. [30]

A comprehensive literature review revealed that there was no 'one' overarching technique or methodological approach to addressing the issue of fragmentation. Instead, the literature identified select informatics research techniques, as used by researchers within the field, which explicitly addressed specific areas of fragmentation.

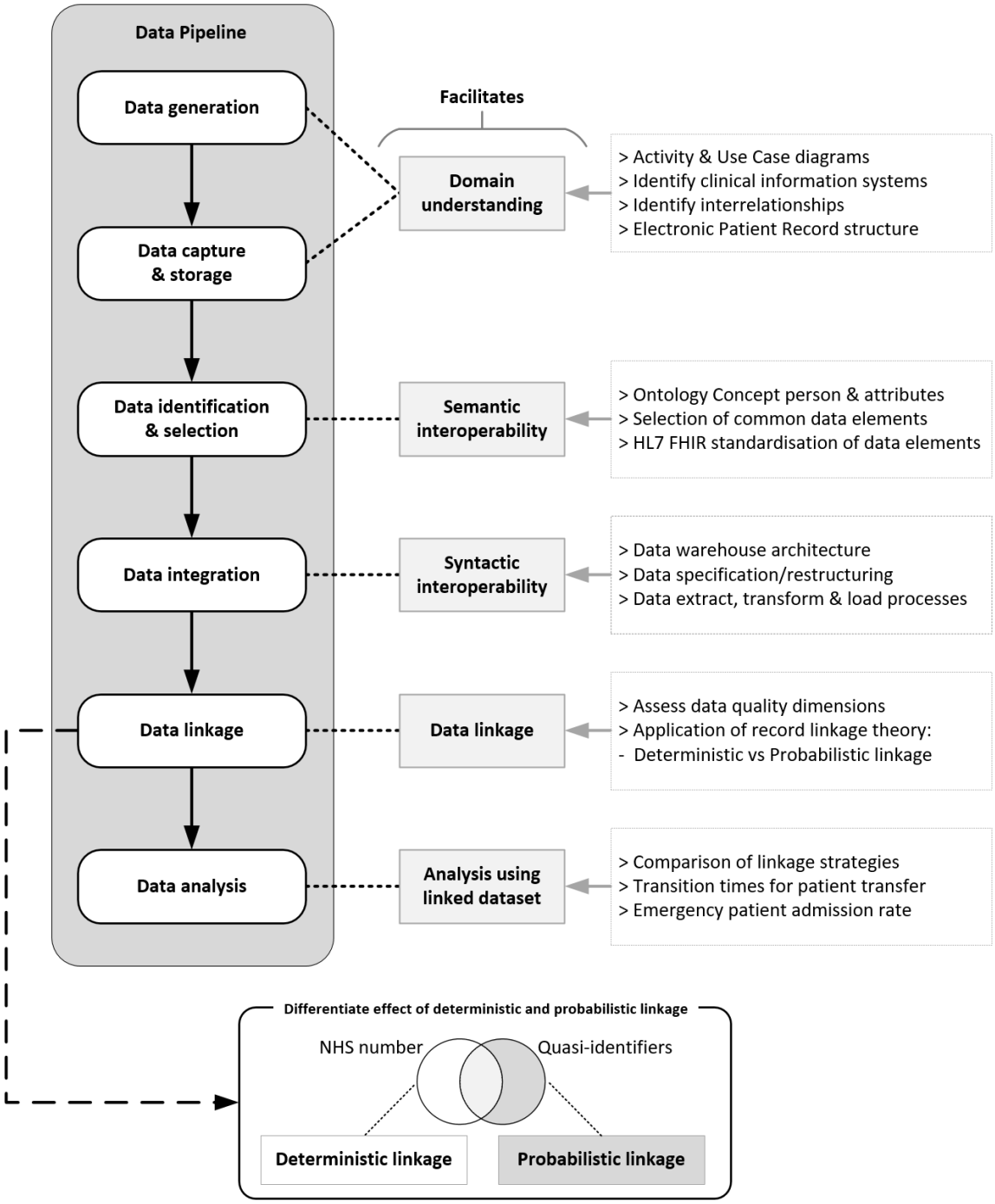


Figure 3.4 Conceptual framework for this thesis

Therefore, as per the research aim, to enable the successful integration and linkage of data silos, the overall approach was subdivided amongst specific tasks, each of which tackled a distinct area of information fragmentation.

Furthermore, GST maintains that the whole of a system is greater than the sum of its parts and emphasises the need for interconnection. [252] In a similar manner, only by using each of the identified techniques in sequential combination will the linkage/integration of data silos be achievable. Thus, a conceptual framework specific for this research was constructed/developed within the context of the research question, to enable the alignment of identified techniques and provide a guide for this research, as illustrated in Figure 3.4 (page 147).

Within the conceptual framework, although the steps may appear distinct, they must be completed in succession, i.e. from the initial to the final step, as each step is informed by, and utilises the output from the preceding step as its input. Additionally, each step encompasses a different task to facilitate domain understanding, semantic/syntactic interoperability and data linkage.

For example, KDD and the data pipeline (page 94) enable domain understanding and provide a means to organise each step. Recognising medical record structures (page 62) aid with CIS data capture and storage, whilst, ontologies (page 82) provide a means to identify common patient attributes amongst the disparate CIS. Additionally, HL7 FHIR (page 91) permits the standardisation of select data elements, and CDW architectures (page 96) enable ETL processes. An assessment of data quality 'fitness' by measuring against dimensions (page 102) permits subsequent data linkage (page 112) by enabling the application/comparison of deterministic linkage (page 113) and probabilistic linkage (page 116) strategies, as per the research hypothesis.

To conclude, the theoretical framework and conceptual framework complement each other, and in the context of this research, the latter was crucial to conceptualising the problem, selecting and linking informatics techniques, structuring the research question and hypothesis by identifying the key aspects important to the phenomenon under study. Additionally, the conceptual framework was crucial for the alignment of the research methodology, as presented in the following section.

3.4 Research methodology

Research methodology is concerned with how to find out about the phenomenon of interest, what approach to use, the principles that underpin it, the type of study to be conducted, and justification for using the adopted research approach. [251] The authors Saunders et al. (2013), [272] use the metaphor of the 'Research Onion' to highlight the key elements that must be considered when formulating/developing an appropriate research methodology. [273] Similarly, working from the outer to the innermost layer, the methodological approach adopted for this research has been outlined in Table 3.2 (page 149).

Table 3.2 Overview of the methodological approach for this research.

Research methodology elements	Approach adopted
Epistemology	Objectivism
Philosophy	Positivism
Approach to theory development	Deduction
Research design	Mono method quantitative
Research strategy	Case study single
Time horizon	Longitudinal

3.4.1 Epistemology

Ontological assumptions about the nature of reality/things, give rise to epistemological assumptions, i.e. ways of researching/enquiring into the nature of reality/things, which influence methodological assumptions, [251] and both inform the theoretical perspective. [273]

Thus, epistemology is the study of knowledge and includes its nature, how it is gained or generated, how it is warranted, and the standards that are used to judge its adequacy. [274] It is concerned with providing philosophical grounding for determining the kinds of knowledge that are possible, whilst ensuring their adequacy and legitimacy, [273] and provides a relation between 'knower' and 'would-be known'. [275]

The philosophical basis for this research is objectivism, whereby, objectivist epistemology maintains that meaning and meaningful reality exist independently from the operation of any consciousness. [273] The author Crotty (1998), [273] provides an excellent example, "the tree in a forest is a tree, regardless of whether is anyone is aware of its existence or not. As an object of that kind

('objectivity') it carries the intrinsic meaning of 'tree-ness'. When human beings recognise it as a tree, they are simply discovering a meaning that has been lying there waiting for them".

In a similar manner, the issue of health information fragmentation is an invisible problem [30] that exists regardless of whether anyone is aware of its existence or not. Although it does not manifest as a physical entity, its effects can certainly be felt. Comparable to the wind, which can neither be seen nor touched, but during a storm, the effects of gale-force winds result in a level of devastation that can clearly be observed and measured. Although the effects of fragmentation may range from subtle to extensive, nonetheless, the problem of 'fragmentation' exists along with what it means to be 'fragmented', and patiently awaits its discovery by humans.

3.4.2 Philosophy

A research paradigm is a way of looking at the world and comprises different philosophical assumptions about what the world is like, how we can understand/know about it, by providing guidance and direction to our way of thinking and subsequent actions. [251, 275] The research paradigm adopted by this study is Positivism. A positivist paradigm rests, partly within an objectivist ontology and a scientific, empirical, hypothesis testing epistemology. [251]

Positivism is based on rationalistic, empiricist philosophy, historically associated with Aristotle, Francis Bacon, John Locke and Auguste Comte. [273, 275] Positivism refers to the importance of what is 'posited', i.e. 'given', which emphasises focus on strictly scientific empiricist method designed to yield pure data and facts uninfluenced by human interpretation or bias. [272]

One of the main assumptions of positivism is empiricism, which maintains that certain kinds of reliable knowledge can only derive from experience. [251] The authors Cohen et al., (2018) [251] state "empiricism means scientifically that the tenability of a theory or hypothesis depends on the nature of the empirical evidence for its support".

Thus, positivists epistemologically focus on discovering observable and measurable facts/regularities and maintain that the use of the scientific method allows experimentation and measurement of what is observable. [272, 273, 275] Hence, only phenomena that can be observed/measured will lead to the production of credible and meaningful data, [272, 273] further supported by general rules/laws to explain any underlying relationships. [275]

Additionally, these universal rules and laws help to explain/predict behaviour and events within organisations. [272] Positivist's state the world exists and is knowable as it really is. [251] Just as physical objects and natural phenomena are real, positivists see organisations and other social entities as real [272], i.e. organisations are real entities with a life of their own. [251]

A positivist researcher will try to remain neutral and detached from the research and data in order to avoid influencing the findings [273] by remaining as 'far as possible', in a value-free way. [272] In addition, positivists may utilise existing theory to develop and test hypotheses, which could be either confirmed or refuted and lead to further development of theory. [272]

Despite its prominence, positivism as an epistemology may not be an appropriate paradigm for all types of research because it relies upon experience as a valid source of knowledge. However, concepts such as space and time are not founded on experience. Furthermore, research findings adopting a positivist viewpoint tend to be descriptive and may lack insight into in-depth issues. [272, 276] Nonetheless, it is the researcher's beliefs/assumptions about the nature of knowledge that ultimately determines their research philosophy.

This research adopts a positivist worldview, as healthcare organisations are considered to be entities with a life of their own, [251] and the phenomena/issue of investigation, i.e. health information fragmentation, exists independently of how it is understood and seen by people. The phenomenon of investigation comprises a set of observable/measurable facts without the researcher 'disturbing' them, i.e. data held securely within CIS that remain unaffected by the researcher. In addition, empirical research permits the identification of patterns, regularities, causes and consequences based on the researcher's experience of information fragmentation, and enables the application of informatics techniques that permit the testing of hypotheses. [269]

3.4.3 Approach to theory development

"Reasoning is the processing and organising of ideas to reach conclusions." [250] Through reasoning, researchers are able to understand and make sense of their thoughts, experiences, and research evidence. [250] Scientific approaches broadly divide into inductive reasoning, which moves from the specific to the general or deductive reasoning, which moves from the general to the specific. [250, 277]

This research takes a deductive approach to theory development, i.e. moving from general premise to a particular situation and conclusion. [250] The deductive approach aligns with logic and simply involves: [276, 277]

$$Theory \rightarrow Hypothesis \rightarrow Observation\ (data\ collection) \rightarrow Finding \rightarrow Hypothesis(confirmation|rejection) \rightarrow Theory\ revision$$

When theorising in a deductive direction, based on what is known about a particular domain, research commences with abstract concepts and outlines logical connections amongst these concepts. Within this context, concepts are abstract ideas that form the building blocks of hypotheses and theories. [278, 279]

A theory about the relationship between these concepts is developed from the academic literature, and a testable proposition/hypothesis is deduced. The hypothesis contains concepts that are translated into researchable entities, and an appropriate research strategy is designed to enable testing of the theory whilst specifying the conditions under which the theory is expected to hold. [249, 277, 278, 280]

Furthermore, appropriate data are collected to measure/analyse the concepts quantitatively and subject the hypothesis to empirical scrutiny and evaluation against evidence for its accuracy under the conditions associated with the particular phenomena being studied. Thus, leading to either a confirmation or rejection of the hypothesis. The findings are interpreted in the context of the predictive ability of the theory, its generalisability and implications for further research. [249, 277, 278, 280]

Therefore, a deductive approach permits the explanation of causal relationships between concepts by measuring them quantitatively and enables the generalisation of the research findings to an extent. [277] Using a deductive approach, this study reviewed the literature focusing on general record linkage theory (page 116). Hence, one could speculate that employing a deterministic strategy would yield different results to employing a probabilistic linkage strategy.

Thus, a hypothesis to empirically test this assumption was deduced, and this study sought to investigate if there was any relationship between the selection and application of a particular linkage strategy on the number of correctly identified records belonging to the same individual across multiple datasets, within a healthcare setting.

The hypothesis for this research (page 21) was stated as:
H_0: *There is no difference between using deterministic or probabilistic record linkage techniques.*
H_1: *There is a difference between using deterministic or probabilistic record linkage techniques.*

Moreover, to facilitate testing of this hypothesis, it was important to utilise a research design, research strategy, and data collection instrument appropriate for this study; detailed in the following sections.

3.4.4 Research design

The quantitative approach toward scientific inquiry transpired from logical positivism, an area of philosophy related to the rules of logic, truth, laws, and predictions. Quantitative research is a formal, objective and systematic process that utilises numerical data to obtain information about the world. [250] The term 'quantitative' is commonly used as a synonym for any data collection technique or analysis procedure (e.g. statistical) that generates or uses numerical data. [280]

Quantitative research designs tend to be associated with a deductive approach, [280] as they focus on using numerical data to investigate relationships amongst variables in order to test theories. [42, 280] Hence, quantitative research enables the testing of theory by describing variables, examining relationships amongst variables and determining interactions (cause and effect) between variables. Quantitative researchers consider 'truth' to be absolute, and that reality may be defined by careful measurement, uninfluenced by personal feelings/perceptions in order to remain objective. [250]

This study adopted a Descriptive Observational Mono Method Quantitative Research design. The purpose of Descriptive research is to describe and interpret the phenomenon of interest (health information fragmentation) as it exists naturally, without interference of any kind from the researcher. [281] Thus, such studies are able to provide a 'picture' of a situation as it occurs naturally. [250]

Observational research focuses on the occurrence of each aspect relating to the phenomenon and quantifies it through some measure, e.g. counts to determine the overall frequency. [281, 282] In addition, Mono method designs utilise a single data collection technique. [280] However, observation studies may require advance planning, meticulous attention to detail, and more time than other descriptive approaches. [281, 282]

Importantly, to enable testing of the hypothesis in relation to underlying theory, [249] this study, informed by academic literature, is designed such that it follows the process outlined in the conceptual framework (page 147), which itself uses pre-established research methodologies, that have been evidenced by their use/application in the design of previous research studies. The replication of earlier research design helps ensure reliability, consistency and reproducibility of this study. [272]

3.4.5 Data collection

Data collection involves a systematic gathering of information in order to address the research questions and hypotheses of a study. [250] Constancy is an essential component of control in the data collection process, whereby it refers to the view that procedures for data collection should present to the consumer as a 'cookbook-like recipe' of how the researcher controlled for the study's conditions. Thus, ensuring environmental conditions, timing, instruments and procedures to collect data are the same for each subject. [249]

Thus, for this research, the instrument to collect data is described within this section, the time horizon is described on page 159, and environmental conditions, i.e. the research setting is described on page 161. The procedures to collect data, i.e. Methods section, are described in detail within the next chapter, page 175.

The utilisation of healthcare data is prerequisite to studying the phenomenon of interest for this research, i.e. health information fragmentation. Therefore, within the context of a mono method quantitative research design, [280] the use of routinely collected hospital administrative data stored within Electronic Patient Records (EPR) can be considered as a single data collection technique.

Healthcare services generate increasingly large amounts of routine data [283] that comprise patient information, diligently collected for the management and treatment of patients. [178, 197, 200] Within a hospital setting, data routinely transcribed within EPRs by clinical and administrative staff [176] contain a wealth of information about a patient including, e.g. demographic, clinical investigations, diagnoses, and prescription details. [283]

Although EPR data are primarily collected to facilitate routine patient care [284], they also have important research potential [283] and support secondary uses for research purposes. However, a researcher engaging in the secondary use of EPR data will have no influence/control over the original

data collection process, which may have taken place many years before. [284] This characteristic of EPR data makes them ideally suited for observational research studies, [285] and align with positivist worldviews, in which researchers are detached, to avoid influencing findings. [273]

An advantage of using routine EPR data is that such data are readily available, useful for identifying hypotheses or undertaking initial assessment and can provide a baseline on levels of health or disease amongst patients. [180] Additionally, such data save researchers time and money, as there are no expensive or logistically challenging data collection, storage, or access requirements. [283, 286] The transcription of data within EPRs is undertaken by 'reliable' and trained personnel, e.g. clinicians and administrators, whose behaviour can be relied upon to be consistent/predictable, this trait provides assurance as to the quality of data being recorded. [249]

Furthermore, routine EPR data enables the creation of a dataset that may incorporate a large patient population spanning several years, in a quick and cost-effective manner. [287] Such data can be stand-alone or potentially linked to other datasets. [285] The linkage of multiple healthcare datasets facilitates the creation of a longitudinal patient pathway, yields a larger sample size (to study rare conditions) and enhances study power to answer research questions. [283]

Moreover, EPRs exhibit stability to record data, measure concepts consistently and provide the 'same' results on repeated administration, over a long period of time, making them ideal for longitudinal studies. [249] Correspondingly, the CIS housing EPRs are maintained and backed up regularly by service providers to prevent any loss of recorded data.

Another attribute is internal consistency or homogeneity, [249] each data element within an EPR is specifically designed to capture one piece of information, e.g. the date/time a patient arrived (YYYY/MM/DD), or patient gender (M or F). Thus, within one CIS, all EPRs will record the same concept using exactly the same format to ensure internal consistency.

However, researchers may face some difficulties when using routine data, as EPRs are heterogeneous and vary in system format, content, software, access/permissions, functionality, [283, 285] and data held may not always be up-to-date. [180] Additionally, there are underlying/hidden issues of EPRs relating to the quality of data, e.g. accuracy, completeness, comparability, that must be addressed [284] using consistent, validated methods. [285]

Nevertheless, EPR data must be of sufficient quality to support any conclusions that result from their use. [284]

Moreover, when using EPRs, there are some important factors to consider, researchers must obtain permission to access the data (security and information governance) and assess data availability and appropriateness for their population of interest. Other factors include data quality assessment, data management throughout the duration of the study, the potential for linking data from different sources. [284, 285] However, due to the heterogeneity of database systems, researchers may not always possess a technical understanding of system nuances, limitations, data models and structures required to access the data, highlighting a requirement for additional training. [285]

Nonetheless, the quantity and real-world nature of routine data provide a strong impetus for their use, [178] as positivist researchers place emphasis on scientific empiricist methods designed to yield pure data uninfluenced by human interpretation or bias. [272] For this research, the conceptual framework (page 147) outlined the methodologies for informatics techniques applicable to EPR data. For example, the importance of assessing data quality was explained in the literature review (page 106) and formed a crucial step within the conceptual framework. Further details about routine EPR data and the clinical databases specifically used within this research are provided on page 172 of this chapter.

3.4.6 Ethics

Ethical research is essential for generating sound empirical knowledge for evidence-based practice. [250] Research ethics are the principles governing how researchers should undertake their work. [249] As per guidance from the NHS Health Research Authority [288] ethical approval was not required for this study as there was no patient involvement and no intervention to patient care.

Thus, this study was a Service Evaluation designed and conducted solely to investigate approaches to the linkage of retrospective routinely collected administrative data within a single NHS organisation, to enable quality improvement. Throughout the duration of the study, the author was employed as an Information Analyst for the Department of Urgent Care and Emergency Medicine at Imperial College Healthcare NHS Trust. [289] In addition, as per hospital requirements, the author had completed relevant Data Protection and Information Governance training, and received specialist departmental Clinical Information Systems training to support the study.

3.4.7 Research strategy

One of the main assumptions of positivism is empiricism, which maintains that certain kinds of reliable knowledge can only derive from experience. [251] As this study adopts a Descriptive Observational Quantitative Research design to study the phenomenon of interest, i.e. health information fragmentation using EPR data, positivist researchers using a deductive approach can use Case Study strategies. [272] The author Yin (2018), [290] defines a case study as "an empirical method that investigates a contemporary phenomenon (the 'case') in-depth and within its real-world context."

Fundamentally, the 'case' forming the basis of an investigation tends to be something that already exists, i.e. it is not an artificially generated situation designed specifically for research purposes. A 'case' needs to be a self-contained entity with distinct boundaries, e.g. types of cases may include organisations such as commercial business, hospitals, schools etc. Furthermore, it is essential that the topic to be studied possess a distinct identity, which permits its study in isolation from its context. [291]

A distinguishing feature of a case study strategy is that it enables researchers to focus on one particular instance of a topic/phenomenon [291] within the context of a real-life setting, to promote in-depth inquiry and understanding. [272, 290] In-depth inquiries can be designed to identify what is happening and why and provide a greater understanding of the effects of a situation, with potential implications for action. [272]

Moreover, the concentration of efforts on one case is beneficial as this yields in-depth insights that may have remained hidden/undiscovered if a different research strategy encompassing many cases had been employed. Hence, the aim being to "illuminate the general by looking at the particular". [291] Thus, case study designs view the 'case' in its entirety, and take a holistic view of what is going on, by appreciating that relationships/processes amongst facets within social settings are interconnected, and how the many parts affect one another. [291]

Strengths of case study research include conceptual validity, via identifying concepts of greatest interest and moving towards their meaning in abstract theory; Being highly heuristic, permitting learning, discovery, problem-solving, new theory construction, development of concepts, and exploring boundaries amongst related concepts. Additionally, capturing complexity by depicting

highly complex, multiple-factor events and tracing processes over time and space; Adjusting the calibration of abstract concepts according to actual experiences and standards; Permitting the incorporation of multiple perspectives and data sources. [272, 278, 291]

Conversely, one of the main criticisms of case study as a research strategy, relates to the generalisability and credibility of their findings and theoretical contributions to knowledge. [272, 291] The viewpoint that findings from case study research cannot be applied beyond the cases studied, is mostly positivist criticisms based on statistical generalisation, which rely upon the use of representative random samples in order to extrapolate findings to a larger general population. [272, 292]

However, gaining a strong understanding of data via in-depth exploration from just one case may identify a wealth of transferable information/knowledge applicable to other similar settings. Thus, shifting the focus of the case study from generalisability to transferability. [292]

Other limitations of case studies include the definition of clear-cut boundaries, which poses difficulties in the selection of appropriate data sources. Ethical considerations may affect researchers gaining and maintaining access to study settings and associated data. Case studies tend to focus on processes that use qualitative or 'soft' data and interpretive methods, rather than measurable outcomes using quantitative data and statistical procedures. [291] Researches may need to demonstrate the extent to which their case study contrasts with other similar case studies. [272]

Despite this, case study strategies have the potential for generating insights from in-depth and meticulous research into the study of a phenomenon in its real-life context, leading to detailed, empirical descriptions and development of theory. [272]

Furthermore, when studying critical, extreme or unique types of cases, researchers may intentionally adopt a single case study strategy, because this provides the opportunity to observe and analyse a relatively uncommon or exceptional phenomenon. [272] Typically, a single case study utilises observational data [275] to explain how and why something happens, by looking in detail at the inner workings of the case. [269]

Some rationales for using a single case strategy include, first, the common case, in which the objective is to capture the circumstances and conditions of an everyday situation, because of the

lessons it may provide about social processes related to some theoretical interest. Second, the critical case, whereby, the theory will have specified a clear set of circumstances within which its propositions are believed to be true. Researchers can then use a single case study to test these propositions. Third, revelatory case, for situations where the researcher has an opportunity to observe and analyse a phenomenon that was previously challenging or inaccessible for study. [290]

3.4.8 Importance of undertaking this research

Moreover, this research investigates the phenomenon of health information fragmentation using a single case study strategy. As mentioned previously, Emergency Medicine is a sociotechnical process [262] in which all actors are interconnected and interdependent on each other to perpetuate action. [32] The Interactive Sociotechnical Analysis model, by the authors Harrison et al. (2007) [31] (page 141) provides an excellent example of the unintended consequences of introducing new Health Information Technologies (HIT) within a healthcare setting.

Introducing new HIT results in changes to the existing social system, and interaction between new HIT with existing technical/physical infrastructures in the environment affect HIT-in-use. [31, 32] The lack of technical interoperability results in the fragmentation of healthcare information and leads to the formation of data silos.

Nevertheless, health information fragmentation remains a widespread problem within healthcare settings. Thus, the single case study presented the author with a unique opportunity to study in-depth the unintended effect of information fragmentation and contributory factors leading to the formation of heterogeneous data silos within one healthcare setting.

This study was based on the Urgent and Emergency Care Services of two large metropolitan teaching hospitals which comprise part of Imperial College Healthcare NHS Trust, London [289]: Charing Cross Hospital (CXH) which has about 500 beds, and Hammersmith Hospital (HH) with around 400 beds. The hospitals provide secondary and tertiary care to inpatients and outpatients and offer a wide range of clinical specialities. [293]

The phenomenon of interest, i.e. health information fragmentation materialised due to the implementation of GP led Urgent Care Centre services (UCC) physically co-locating with EDs at both HH and CXH sites, with primary care GPs acting as gatekeepers. Thus, a proportion of patients

attending the UCC required treatment within ED, of which some were admitted to hospital Inpatient Care (IPC) services. Each respective service utilised a bespoke CIS to capture EPR data; however, these systems were not interoperable and did not 'talk' to one another, resulting in the formation of data silos. This naturally occurring phenomenon provided an excellent opportunity to undertake research via single case study strategy. Further information about the research setting is provided on page 161 of this thesis.

Additionally, when considering the time factor for a study, the main strengths of longitudinal research is its capacity to observe the same phenomenon over an extended time frame and also for observing change and development over time. [272, 294] EPRs are ideal for such studies as the linkage of multiple healthcare datasets enables the creation of a longitudinal patient pathway. [283]

This study utilised routinely collected EPR data from three department UCC, ED and IPC.
As the data had been transcribed historically, the phenomenon of interest was being studied retrospectively. The UCC services were implemented in 2009; hence, this study employed a longitudinal design with the study period spanning seven years, from 01st January 2007 to 31st December 2013.

Thus, permitting an in-depth study of information fragmentation and reconciliation of patients attending urgent and emergency care services more than once. Additionally, it enabled the creation of a Linked dataset that enabled calculation of the transition times for patients transferred between services (research objective seven), reconciliation of the four-hour standard using the Linked dataset (research objective eight) and a comparison of the rate of emergency patient admissions pre/post UCC implementation (research objective nine).

To conclude, this study enabled the development of a conceptual framework (page 147) to align and apply informatics techniques and utilise record linkage theory to facilitate the linkage/integration of disparate data silos, by employing a descriptive, observational, mono method quantitative, single case study, longitudinal research design. The methods chapter (page 175) provides specific details about the approach used in order to address the research questions and objectives of this study.

3.5 Research setting

3.5.1 The London Borough of Hammersmith & Fulham

Hammersmith & Fulham (H&F) situated in the centre-west of London, is the country's sixth most densely populated area [295] and is the third smallest of the 13 inner London Boroughs. H&F lies within the top 50 most deprived local authority areas in England, ranking 31 from 326 Local Authorities, and 13 out of the 33 London Boroughs, based on the Index of Multiple Deprivation (IMD). [296, 297]

The IMD is a national index developed by the UK Government to measure a range of indicators including Income; Employment; Health & disability; Education, skills & training; Barriers to housing & services; Living environment and Crime. [297]

The population of H&F has risen by approximately 12 per cent, from 165,242 in 2001 to 185,004 in 2018, [296] and the age profile is typical of an affluent urban population. Three out of four residents are within working age (16 to 64 years), and there are fewer people near retirement age and a correspondingly lower level of younger children than in London and England & Wales. [295]

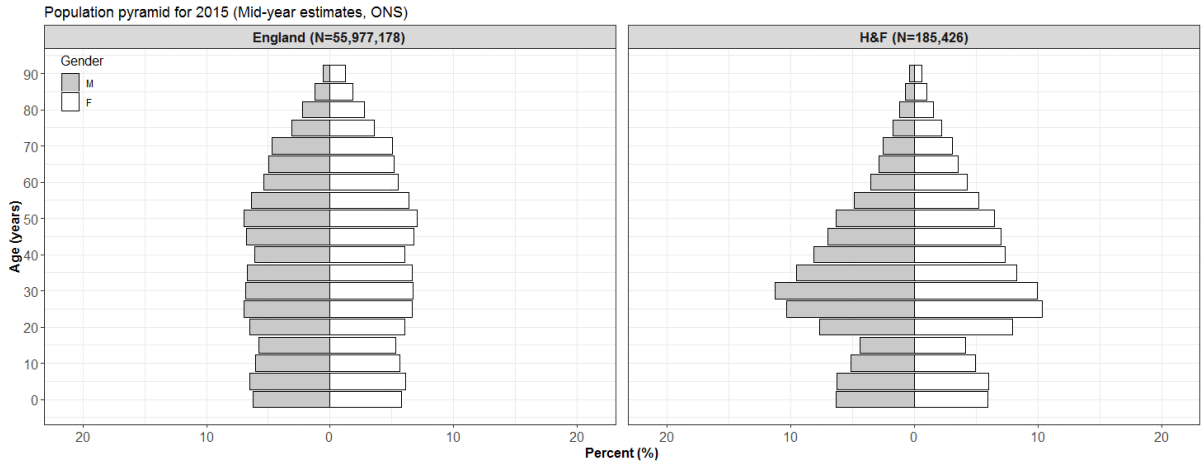


Figure 3.5 Population pyramid of England and Hammersmith & Fulham, 2015 (source Office of National Statistics).

Figure 3.5 (page 161) adapted from the Borough profile H&F Council (2018) [296] compares the proportion of residents in England and H&F, by age and gender. Compared to England, H&F has a higher proportion of young adults aged between 25 to 40 years, which is typical of an affluent urban population. The proportion of females (51.3 per cent) is marginally higher than males (48.7 per

cent), and 58 per cent of residents belong to White British ethnic groups. Approximately 86 per cent of H&F residents regard themselves as having 'good' or 'very good health'. [295]

3.5.2 Healthcare services within H&F

The H&F Borough accommodates 30 GP Practices and three NHS hospitals, one of which operates across the neighbouring Boroughs of Chelsea & Westminster and Borough of Kensington. The remaining two hospitals are geographically located within H&F Borough and comprise, first, Charing Cross Hospital (CXH), which is a general acute hospital providing a range of adult clinical specialties, housing the serious injuries centre for West London and is a key site for the teaching of medical students from Imperial College London. [298]

Second, Hammersmith Hospital (HH) which encompasses the West London Renal and Transplant Centre, two large cancer treatment centres, specialist cardiac services with a 24/7 primary angioplasty centre and the largest Clinical Imaging Centre in Europe. [298]

Both hospitals, CXH and HH, fall under the umbrella organisation Imperial College Healthcare NHS Trust (ICHNT), formally recognised on 1st October 2007. ICHNT is one of the largest Trusts in England, and provides acute and specialist healthcare, serving a population of nearly two million people in North West London. It is comprised of five hospitals: Charing Cross, Hammersmith, Queen Charlotte's & Chelsea, St Mary's and The Western Eye. These five hospitals integrated with Imperial College London to form one of the UK's seven Academic Health Science Centres. [289, 299]

3.5.3 Urgent and Emergency Care

In order to facilitate an understanding of the phenomenon of interest, i.e. information fragmentation and subsequent formation of data silos, this research focuses specifically on the EDs at both CXH and HH sites. Analogous to national trends (page 37), the resident population of H&F Borough is increasing, which has inevitably resulted in a rising number of emergency attendances and consequent admissions at both hospital sites.

To tackle these issues, an integrated urgent care model commissioned by NHS Hammersmith and Fulham in 2009, led to the establishment of two Type 3 (page 29) 'General Practitioner-led Urgent Care Centres' (UCC) co-locating with Type 1 Major EDs at both CXH and HH sites. This 'Partnership

for Health' was a collaboration between ICHNT, NHS Hammersmith & Fulham Commissioning Care Groups and London Central & West Unscheduled Care Collaborative (out-of-hours provider).[34, 300-302]

The ED at CXH site treats on average 70,400 patients a year (193 per day), and its UCC is operative 24 hours a day, seven days a week. The ED at HH site is comparatively smaller and treats 31,300 patients a year (86 per day), and its UCC is operative between the times 0800 and 2200, seven days a week. Comparatively, traditional GP core hours are between 08:00 and 18:30 Monday to Friday. [37, 40]

The integrated care model was based on 'Model three', primary care front-end, with GPs screening/filtering and treating the majority of 'walk-in' patients or self-referrals to ED [66, 92] (page 50), whereby, the UCC formed the front-end of the ED, with GPs acting as gatekeepers.

This delineation of services facilitated clarity of responsibility, and following consultation and preliminary piloting, GPs and hospital specialists jointly developed novel streaming guidelines, prior to the implementation of the integrated urgent care model at both HH and CXH sites. To factor in time for adjustment, the UCC's launched in phased stages, UCC at HH commenced April 2009, and the UCC at CXH commenced September 2009.

Additionally, the 'Partnership for Health' collaboration also provides a completely separate primary care service, based on 'Model two' (page 50), the location of GP service alongside ED, that operates in parallel to the UCC, and is physically located alongside the ED. However, this service works in complete isolation and operates independently from both the UCC and ED services, and was therefore not included in this research.

Pre-UCC implementation

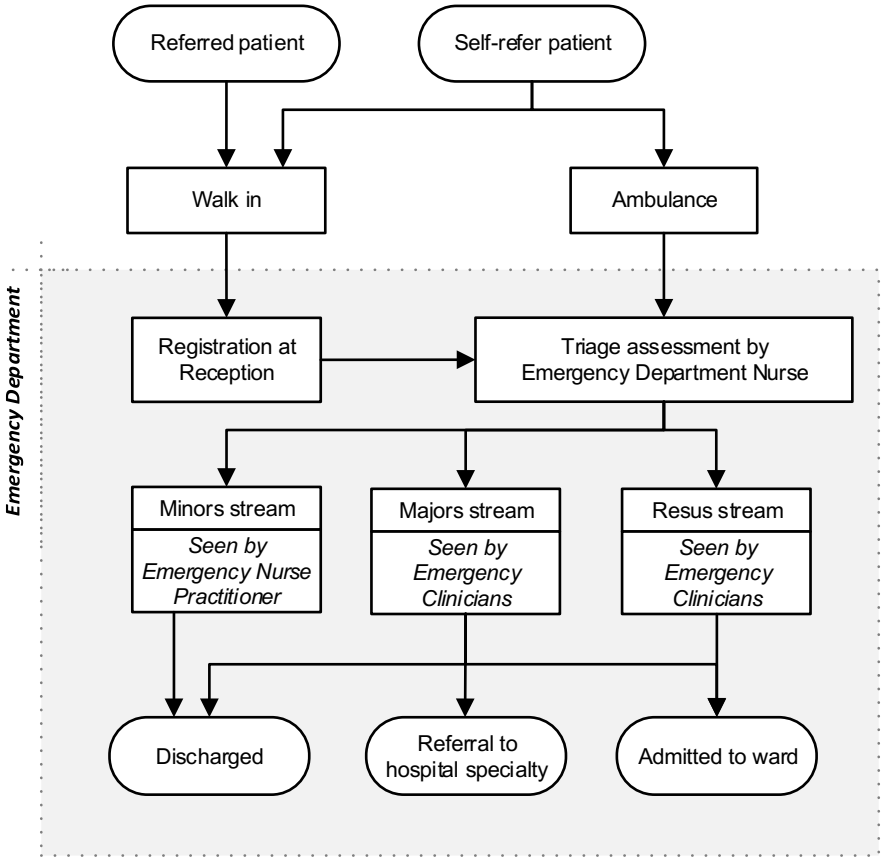


Figure 3.6 Emergency Department pathway pre-UCC

Prior to the implementation of the UCC, Figure 3.6 (page 164) provides an overview of the operational pathway for ED, from patient arrival to disposal outcome. Patients may arrive via walk-in, referral from another healthcare provider or ambulance, present to ED and are assessed by a triage nurse, who based on the patients' condition, selects a triage category (page 31) and allocates a stream.

Patients allocated to the Minors stream are seen by an Emergency Nurse Practitioner (ENP) and discharged home, other patients would be allocated to either the Majors or Resuscitation streams to be seen/treated by ED clinicians, with outcomes that include being discharged home, referred to specialty or admitted to hospital.

Post-UCC implementation

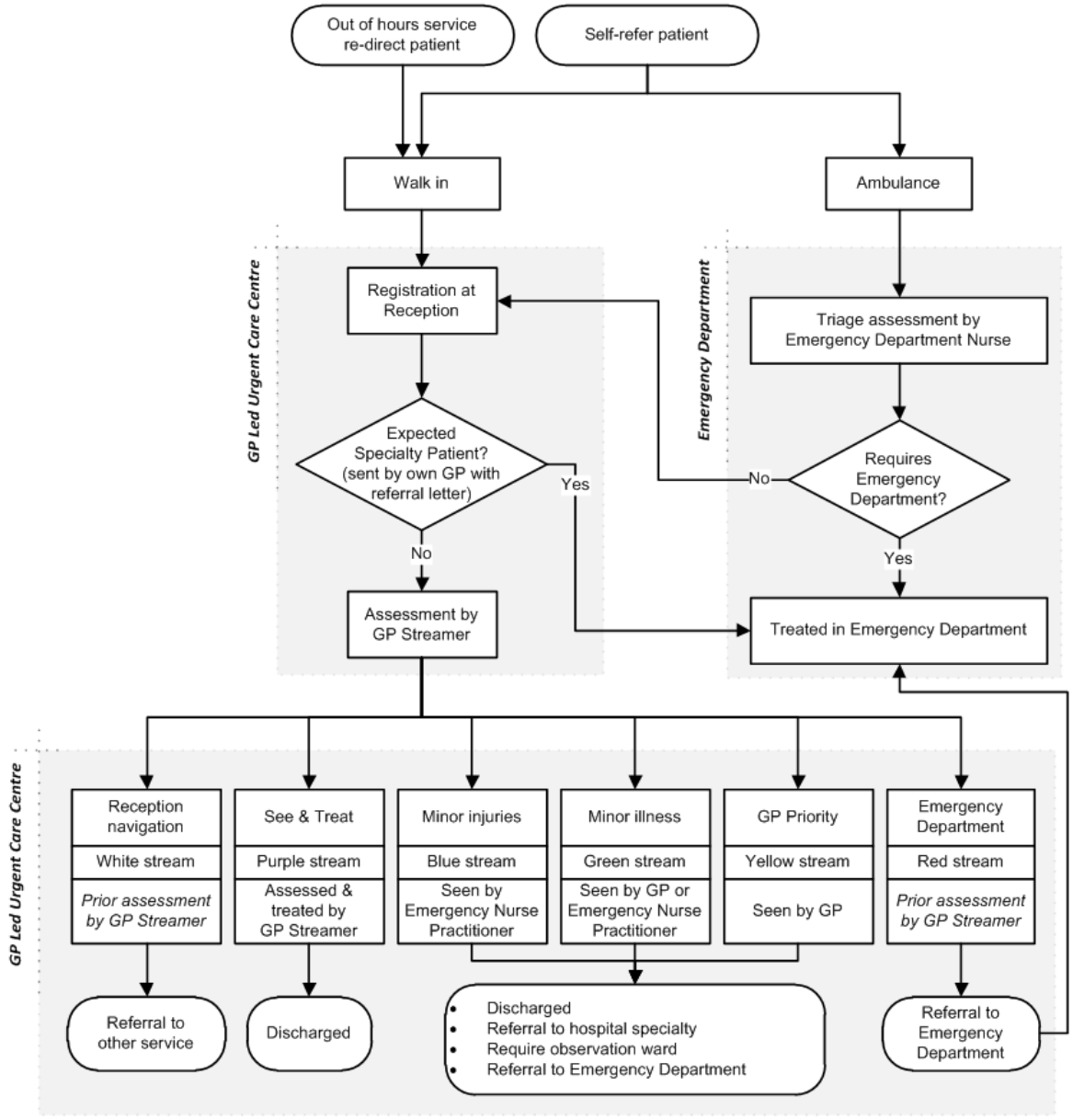


Figure 3.7 Patient pathway across the integrated GP-led model of care (Gnani et al., 2013)

The operational pathway Figure 3.7 (page 165) [34] illustrates the updated integrated model resulting from the inclusion of the UCC. Following its implementation, the dynamics at both EDs at CXH and HH changed clinically, operationally and physically. As the entrance to the UCC, waiting area and environment generally remained the same, incorporating some new signage, patients would perceive they were still visiting the ED.

In this new model, all 'walk-in' patients (i.e. patients not arriving by ambulance) reported to the UCC reception and were streamed by GPs, and those patients with primary care type conditions were subsequently treated within the UCC by GPs or ENPs. Only patients, whose conditions the GP

streamer felt necessitated emergency care were directed to ED as appropriate; this enabled ED staff to focus on higher-need patients that arrived via ambulance or were referred from the UCC. [34] As such, the model was designed to aid in the reduction of issues related to overcrowding, allowing patients to be seen and treated more efficiently, alleviating the number of patients requiring emergency admission and consequently helping to improve the four-hour standard.

From a clinical perspective, the integrated model enabled ED clinicians to focus their attention and resources on patients with critically ill or life-threatening conditions, whereas GPs in the UCC concentrated on treating those patients presenting with relatively minor injuries or illnesses.

However, in some instances, patients that had arrived by ambulance and following assessment by the ED triage nurse were found to have primary care type conditions, were subsequently referred to the UCC for treatment. Table 3.3 (page 166) provides a brief description of the streams illustrated in Figure 3.7 (page 165).

Table 3.3 Overview of streaming designations (adapted from Gnani et al., 2013)

Designation	Description	Colour
Emergency department	Patient seen in the emergency department	Red stream
Expected special patients	Patient arrives with letter from own GP	Red stream
GP priority case	Patient seen by UCC GP	Yellow stream
Minor injury	Patient seen by nurse practitioner	Blue stream
Minor illness	Patient seen by GP or nurse practitioner	Green stream
Reception navigation	Patients signposted to other services	White stream
See and treat	Patient treated at the point of streaming	Purple stream
Did not wait	Patient did not wait to be seen	Black stream

Furthermore, to ensure that patients were reviewed by the most appropriate clinician the first time, streaming guidelines were developed by the collaboration of two GPs, two Emergency Medicine Consultants and two Emergency Nurse Practitioners. The guidelines considered the 'type' of patients that would normally walk through the door of the ED and the case-mix UCC GPs would be happy to see, in the context of normal primary care presentations. Cases that were exceptions to these rules would not be seen by GPs, and such patients were assigned to the 'Red stream' to be seen within the ED. [34]

HH site patient attendance and admissions

Figure 3.8 (page 167) provides an overview of the number of episodes of attendance and admission per month throughout the study period, pre/post UCC service implementation at the HH site. Pre-UCC implementation the trend for attendances remained consistent until month 25 where attendances begin to rise. However, post-transition month (not included) the overall trend continues to increase steadily for a number of years, reaching the highest point (month 65) and appears to decrease marginally thereafter, almost reaching a plateau. Conversely, upon visual inspection, the number of monthly admissions pre/post UCC appear to be consistent to month 64 after which a level change is discernible.

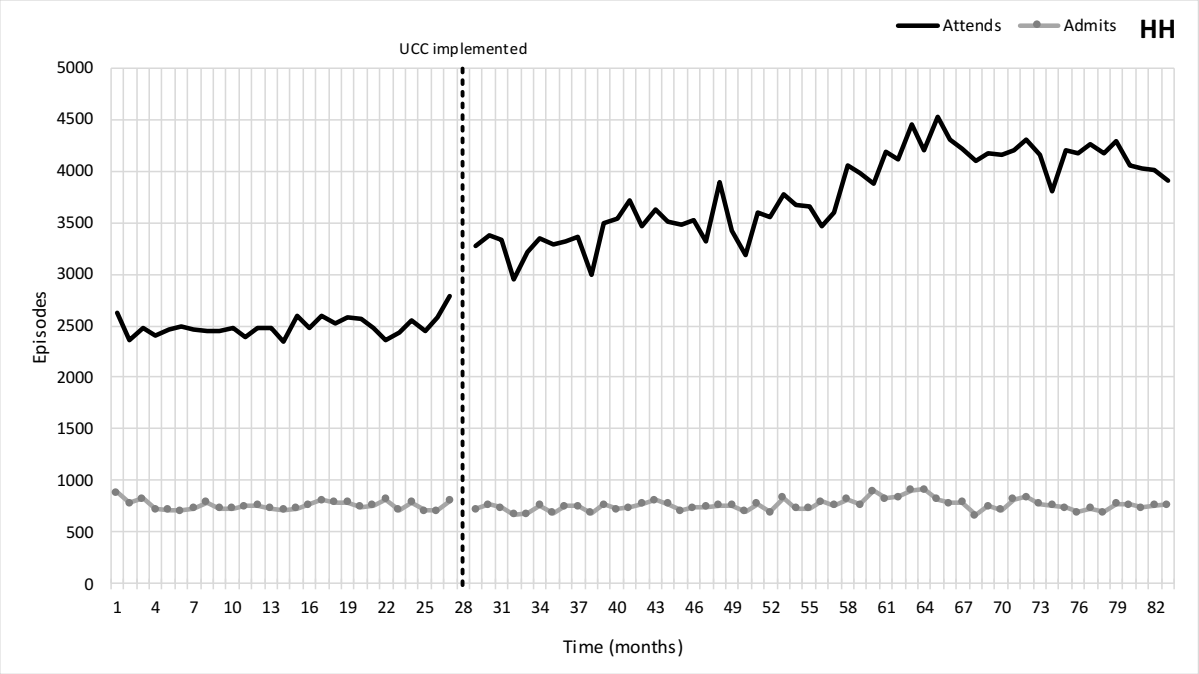


Figure 3.8 Episodes of attendance and admissions throughout the study period at HH site

CXH site patient attendance and admissions

Figure 3.9 (page 168) provides an overview of the number of episodes of attendance and admission per month throughout the study period, pre/post UCC service implementation at the CXH site. The trend for attendances appears to increase throughout the study period, remaining between 5000 to 6000 attendances per month pre-UCC. However, from month 38 to 39 there was a sharp increase in the number of attendances, pushing the number of attendances from 6000 to 7000 attendances per month. Similarly, the number of monthly admissions pre-UCC appear to remain at an almost consistent level. However, post-UCC the number of monthly admissions begin to increase gradually, in line with monthly attendance.

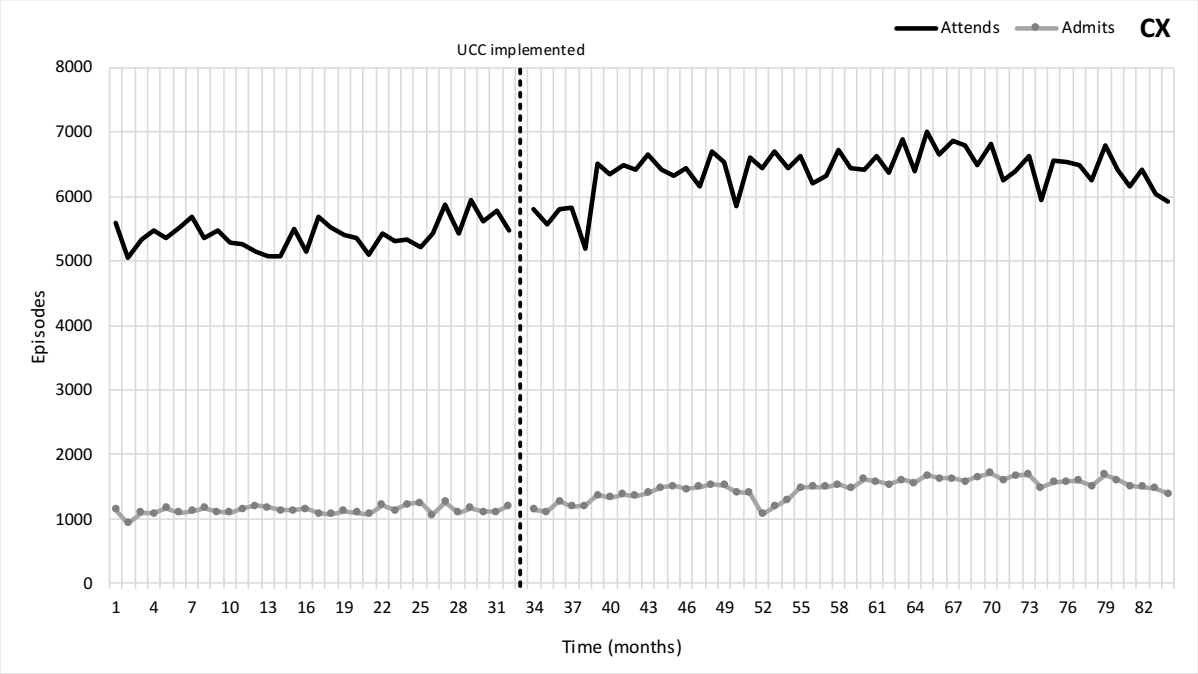


Figure 3.9 Episodes of attendance and admissions throughout the study period at CXH site

ICHNT Urgent and Emergency care staffing

To deliver timely and high-quality care within the UCC and ED, correct staffing, taking into account capacity and capability is crucial. However, different departments will have different needs, and there is no 'one-size-fits-all' staffing model. [303]

Over time, the workload for EM staff has gradually increased, due to the nature of illness and complexity of conditions for patients presenting to ED and seasonal effect. This has meant that more had to be done by ED healthcare personnel, i.e. patients being assessed and fully worked up; taking a patients' history; undertaking procedures and tests to reach a clinical diagnosis before the patient would be either discharged or an inpatient admission arranged. However, decisions to admit patients may be indirectly affected by physical space available, as hosspital buildings accommodate different numbers of beds.

Thus, staffing levels have changed appropriately to reflect this; for example, in 1982, there was only 1 Consultant, 0 Middle grades, 0 Registrars and 7 SHOs. Table 3.4 (page 169) provides an overview of the type and number of staff, per site, pre/post UCC implementation. At the HH site, pre-UCC there were neither GPs nor ENPs, only ED staff, however, post-UCC there were zero ED staff, and only GP and ENPs.

Conversely, at the CXH site, post-UCC saw an increase in the number of ED staff, and the Minors Unit which employed one GP (who did not triage) was closed when the UCC commenced. Post-UCC, both CXH and HH sites, utilised more ENP hours, which is reflective of the number of minors-type patients that present.

Table 3.4 Staffing per site pre/post UCC

Staffing at HH (approximate figures)	**Pre-UCC (Jan 2007 to Mar 2009)**	**Post-UCC (Apr 2009 to Dec 2013)**
Emergency department		
Consultants	1	0
Registrars	1	0
SHO's	5	0
HH UCC (over 24 hours)		
GP	---------	29 hours (4 shifts)
ENP	---------	40.5 hours (4 shifts)
Staffing at CXH (approximate figures)	**Pre-UCC (Jan 2007 to Sep 2009)**	**Post-UCC (Oct 2009 to Dec 2013)**
Emergency department		
Consultants	3	6
Registrars	5	6 or 7
SHO's	8	11
Minors Unit		
Sessional GPs 1 employed F/T by Trust	2 x 4 hour sessions	---------
ENP	4	---------
CXH UCC (over 24 hours)		
GP	---------	40 hours (5 shifts)
ENP	---------	50 hours (5 shifts)

Population age/gender pyramids

The patient population pyramids for HH (page 170) and CXH (page 171) serve to illustrate the distribution of patients that present to ED, UCC and IPC, before and after UCC implementation.

HH site

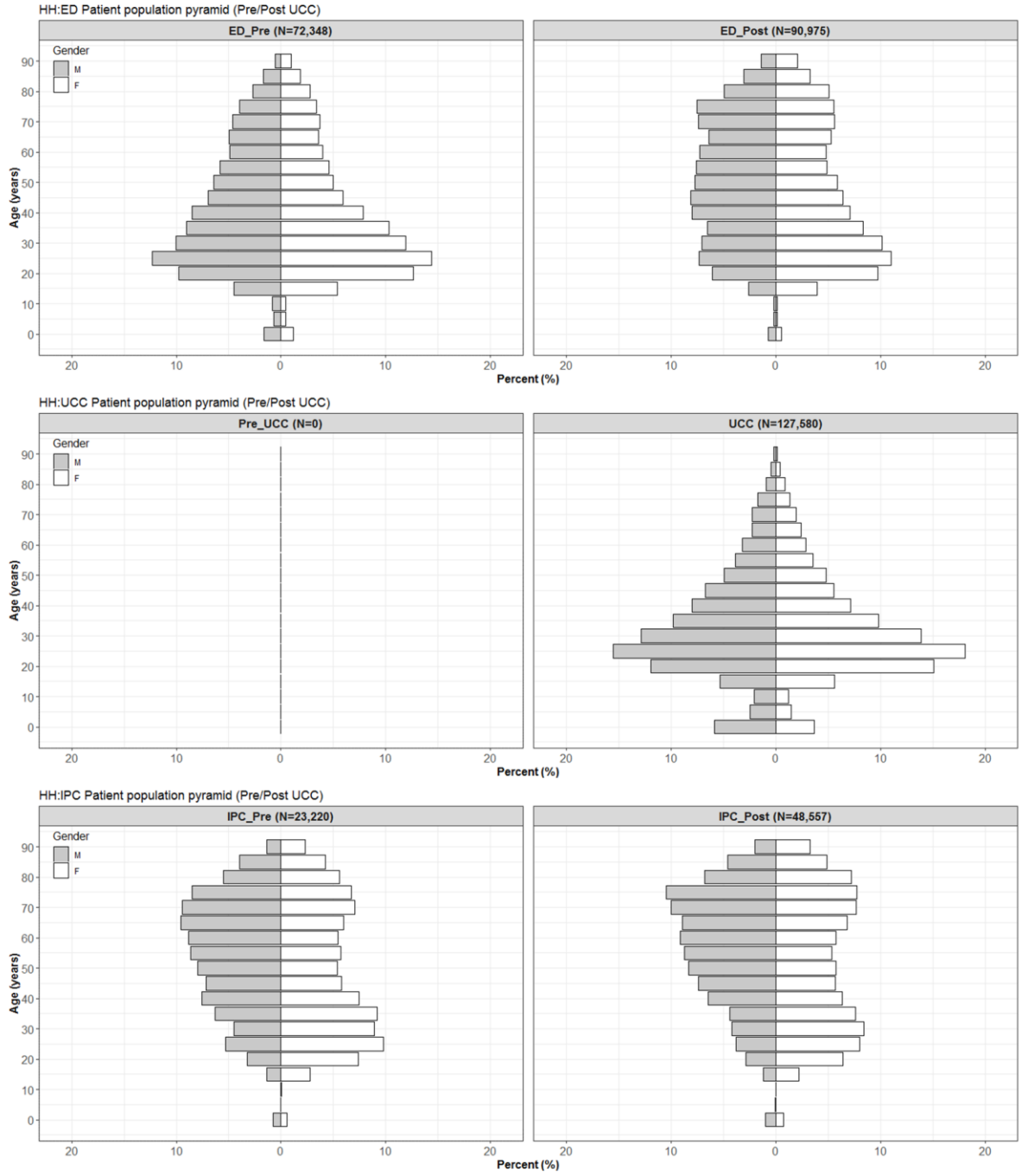


Figure 3.10 HH site patient population pyramid pre/post UCC

CXH site

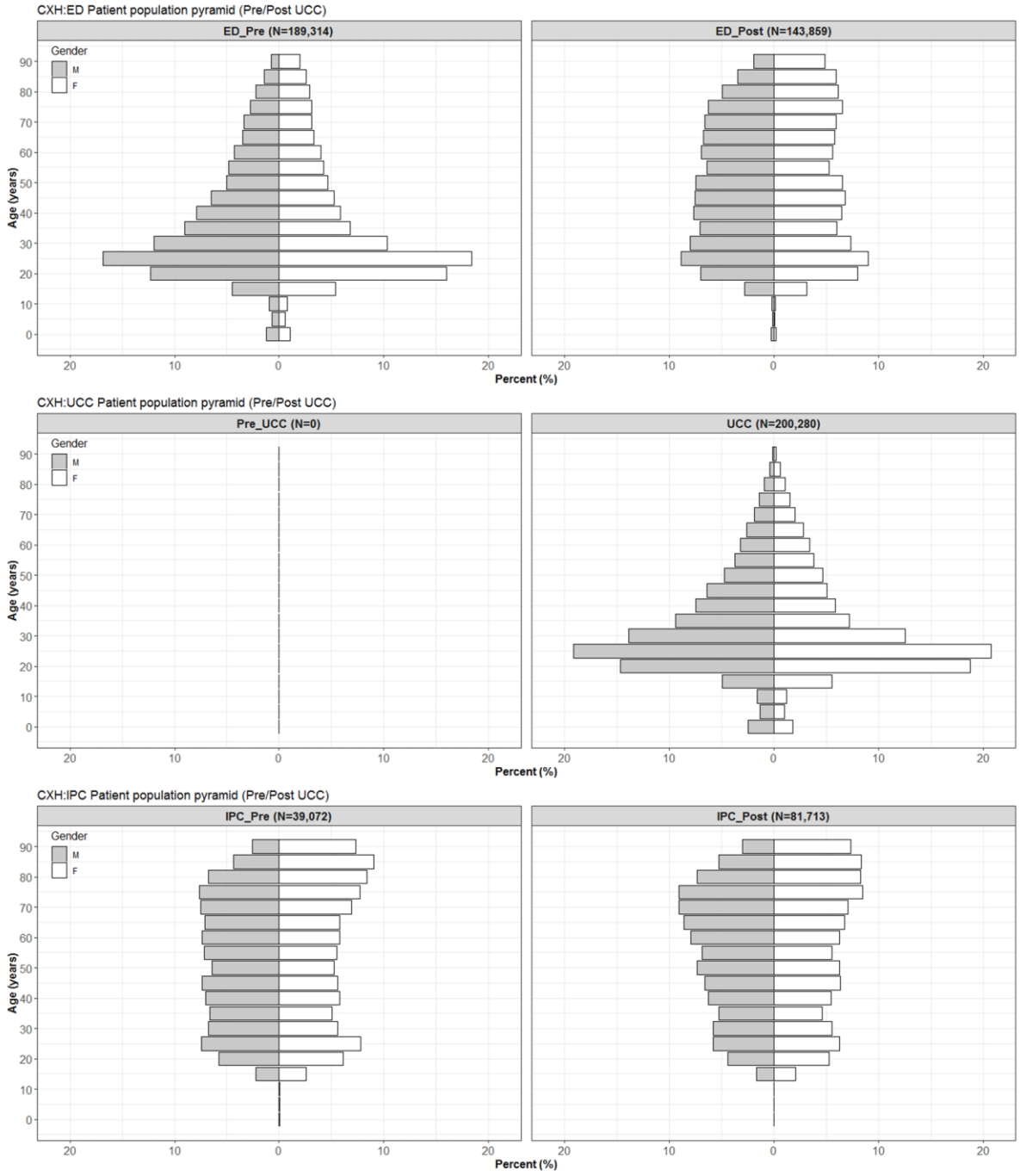


Figure 3.11 CXH site patient population pyramid pre/post UCC

Pre-UCC

Both HH and CXH sites show a large proportion of young adults aged between 20 to 40 years attend ED, which is as expected when compared to the pyramids for England and H&F Borough, Figure 3.5 (page 161). Additionally, there are very small proportions of children aged under 10 years attending either site, and the percentage of female patients is slightly higher than males. Referring to the charts for patients admitted to IPC, HH has a larger proportion of young female admissions than CXH, possibly due to the same site location of Queen Charlotte's & Chelsea Hospital, which specialises in maternity, women's and neonatal care.

Post-UCC

The introduction of the UCC at both HH and CXH sites resulted in a substantial observable effect on the number of patients seen in ED, as the majority of patients attending the UCC are young adults. The charts for ED at both CXH and HH sites appear similar, but with slightly more elderly admissions at CXH. Referring to IPC, there was a slight reduction in the number of young adult admissions at HH but marginal change at CXH site.

3.5.4 Clinical Information Systems

At the time of writing this thesis, ICHNT contained 750 official databases, maintained by specific clinical and non-clinical departments within the Trust. One-third of these were stand-alone, and this figure excludes any additional unknown 'in-house' departmental data systems. [304]

Within the context of this research, urgent and emergency care services encompass the UCC, ED and IPC settings, and contain three distinct CIS respectively, Figure 3.12 (page 173) illustrates the time when each CIS was implemented at each hospital site, and Table 3.5 (page 173) provides an overview of each CIS.

The IPC database was the oldest of the three CIS, followed by the ED database, with the most recent being the UCC database. Furthermore, system design, software platform and technologies used within each CIS will reflect the technology available at that moment in time, resulting in a natural divide between legacy and modern systems.

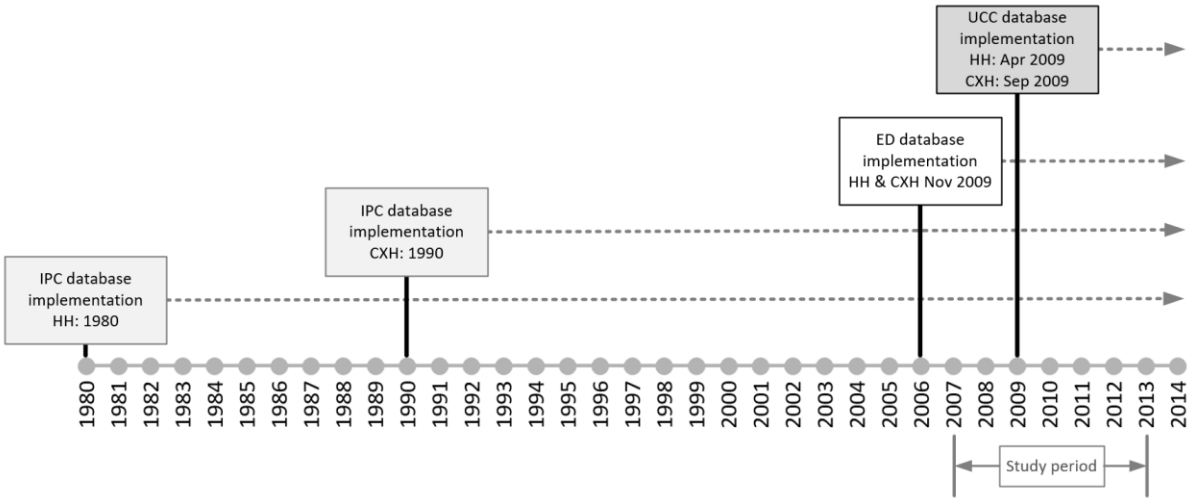


Figure 3.12 Timeline of departmental clinical information system implementation.

The IPC databases contained information about emergency patient admissions, whilst the ED and UCC databases contained information about emergency patient attendance. Thus, for any patient that required treatment within these three departments, their entire 'episode of care' was being transcribed in 'pieces' or fragments within each of these three databases.

Table 3.5 Overview of departmental Clinical Information Systems

	Department/Service		
CIS detail	**UCC** (GP Urgent Care Centre)	**ED** (Emergency Department)	**IPC** (Inpatient Care)
System name	Adastra	Symphony	ICHIS
Department	UCC only	ED only	Main hospital system
Data stored	Patient attendance	Patient attendance	Patient admissions
Implemented	HH: Apr 2009 CXH: Sep 2009	HH & CXH: Nov 2006	HH: 1980 CXH: 1990
Users	Clinical & Administrative	Clinical & Administrative	Clinical & Administrative
Used for	Patient management Financial Administrative Clinical audit Mandatory reporting	Patient management Financial Administrative Clinical audit Mandatory reporting	Outpatient & Inpatient Financial Administrative Clinical audit Mandatory reporting Diagnostic tests Admission spells Used by all specialties

Both UCCs at CXH and HH sites use the 'Adastra' clinical patient management system [305], to capture both patient administrative information and clinical data. This system is predominantly used

by other GP out of hours services within the area. [34] In the year 2009, the UCC services implemented using a staged approach, with HH site going live in April 2009, and later that year the UCC service at CXH went live in September 2009, to permit adjustment time for each service.

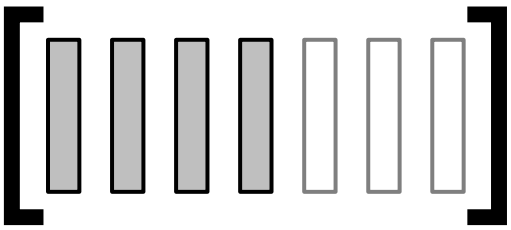
Both ED sites use 'Ascribe Symphony' [306] system, to record clinical and administrative information about a patient's episode of care, implementing simultaneously in November 2006. However, it was not possible for the vendor to import any patient records held within the preceding ED database; hence, Symphony did not contain any historical records pre November 2006. Further investigation revealed that the preceding ED database was taken offline, and all data contained within it was now inaccessible. Furthermore, the Symphony parent company 'Ascribe', was later subsumed by another vendor, EMIS Health, but this did not affect its functionality.

The ICHNT main Patient Administration System (PAS) is 'Integrated Computerised Hospital Information System' (ICHIS). [307] It is the oldest of the three CIS, in continuous use since 1980 at HH site and following the merger with CXH, became the principal PAS covering both sites. All specialties within the Trust use ICHIS to manage patients, and this was accomplished by splitting the system into numerous smaller modules, each with a specific function, e.g. demographic module, diagnostic test result module. Additionally, the ICHIS system captures all details of patients admitted within the IPC module, including emergency admissions via ED.

3.6 Chapter summary

In summary, this section presented the importance and appropriateness of Sociotechnical Systems theory to frame the issue of information fragmentation and developed a conceptual framework to align and apply informatics techniques sequentially. Furthermore, it employed a descriptive, observational, mono method quantitative, single case study, longitudinal research design, to facilitate an in-depth study of information fragmentation and data silos, as faced by the emergency care services at both CXH and HH hospital sites.

Chapter 4 Methods



4.1 Introduction

This chapter presents the methods for this study, aligned to meet the specific objectives of this research and reflect the sequential order of processes within the conceptual framework. Thus, methods that enable Domain understanding, Semantic interoperability, Syntactic interoperability, Data quality, Data linkage and Linked data analysis, are presented henceforth.

4.2 Domain understanding

Addresses objective 1: To explore the formation of data silos within healthcare settings.

4.2.1 Process modelling

In order to meet objective one, this research took a Business Process Modelling (BPM) approach using the Unified Modelling Language (UML). BPM approaches enable the representation of an organisation's processes and workflow in a graphical manner. [308, 309] UML is a visual modelling language primarily used within software engineering and is process independent, so can be used to model a variety of domains, e.g. from businesses to software implementation. [310]

Activity diagrams are a type of UML behaviour diagram and used to model business processes, whereby an 'activity' is a parameterised behaviour, represented as a coordinated flow of action that places emphasis on the sequence and conditions of the flow. [309, 310] Such diagrams incorporate activity states, action states, the transition of data between activities, and sequential, branching/joining transitions. Additionally, activities may be grouped to represent organisational structures and relative execution sequences. [308]

Another type of UML diagram is the 'Use case' or behaviour diagram which describes a set of actions (use cases) that some system (subject) can perform in association with external users of the system (actors). Thus, Use case diagrams represent the totality of the system in the form of entities, processes and their interactions, permitting a snapshot of the organisational and behavioural elements of a system. Moreover, each Use case provides an observable result for each of the 'actors' within the system. [308, 310]

Therefore, to facilitate an understanding what CIS were being utilised and how information was being recorded within the Urgent and Emergency Care setting, the author spoke with clinical and administrative staff and followed the footsteps along a 'virtual' patient pathway that included the

UCC, ED and IPC departments, with results visualised as a UML activity diagram. Additionally, a Use case diagram depicting the various interactions between patients, staff and CISs was created, to provide a different perspective.

Nonetheless, the most critical step to gaining an understanding of the urgent and emergency care setting was to speak with, and obtain the perspectives from, numerous clinical, managerial and administrative staff, to enable modelling of departments and mapping of the diagrams, reiteratively.

Furthermore, a key factor to consider is how much data to collect? Research studies require a certain amount of specific data which must be collected to fulfil its aims. Beyond these data, there are often other data that could also be obtained, but it is important to weigh the costs and consequences of collecting too little or too many data. Too little data results in missed data, and if not collected, may not be available for collection at a later occasion. Conversely, collecting too much data is likely to add to the time and cost to data collection and processing, and may threaten the completeness and/or quality of key data items. Thus, it is essential to decide at the start of any research, what key data are required. [179]

4.2.2 Training

At the heart of the CIS lie patient data, and the configuration in which CISs have been designed and organised is essential to the functioning of the system. [124] A typical database may contain hundreds of tables and requires an associated data dictionary to facilitate navigation and location of specific data elements within the relational model of a database. Data dictionaries are often live documents created by database developers that detail every specification of all data elements within tables, i.e. primary identifiers and underlying relationships. [122]

Therefore, to facilitate greater understanding of each of the CISs used within the UCC, ED and IPC services, the author liaised with Database Administrators, System Vendors, and underwent specialist system training. This process enabled the author to gain an appreciation of the nuances of each CIS, the structure, layout, how/where information was being recorded, and naming conventions used. Furthermore, to understand the underlying relational database models for data extraction purposes, the author was granted access to a copy of the data dictionary for each CIS.

4.3 Semantic interoperability

Addresses objective 3: To apply techniques that enable interoperability between the clinical information systems in the emergency care setting.

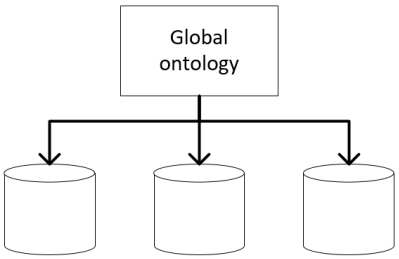
4.3.1 Ontologies

One approach identified in the literature, and applied within the field of Information Systems is to use conceptual modelling, i.e. Ontologies (page 82). [150] An ontology is an explicit specification of a conceptualisation [148] and related concepts can be grouped into ontologies. [149]

Thus, an ontology offers a convenient means of connecting a subject, a relationship, and an object together. Ontologies permit the organisation of concepts under a common specification in order to facilitate knowledge sharing, and therein lies their value.

For example, when considering the abstract concept of a 'Person' via ontologies, data type properties, i.e. First Name, Last Name, Gender, Date of Birth, and Home Address, can be shared with other users and interpreted in the same uniform way. [149] Thus, a single ontology approach [157] can be utilised to create a patient ontology with specific properties common to all three CISs.

Global/single ontology approach
An integrated ontology that describes data in all sources, and all queries via global ontology.



4.3.2 Standardisation of data elements

The data elements within the three disparate CIS were heterogeneous and used different naming conventions for common data elements. In order to facilitate the integration and linkage of data elements, the name, format and coding conventions must be exactly the same across all datasets to enable successful identification and subsequent integration. [179]

The literature identified a standards framework created by HL7, known as Fast Healthcare Interoperability Resources (FHIR) [160], (page 91), and another approach referring to the NHS Data Dictionary, as used by NHS Digital. [311]

These approaches provide guidance on how to standardise the naming conventions of data elements along with the specifications for data types, formats and coding. For example, the data element 'Gender' could also be named 'Sex', and coded as either Male/Female, M/F or 0/1.

Thus, to facilitate the exchange of healthcare-related information, and enable linkage and integration, data elements must have the same meaning semantically across each of the three disparate CIS, which can be accomplished using the aforementioned approaches.

4.4 Syntactic interoperability

Addresses objective 4: To select data warehouse architecture appropriate for housing healthcare data extracts.

As the UCC, ED and IPC databases were vendor-locked, it was not possible to directly access the underlying data tables, nor set up any processes to transfer of information between each CIS as there was no syntactic interoperability.

4.4.1 Clinical Data Warehouse

Within the computing discipline, one approach to integrating data is to follow a specific sequence of steps, envisioned as a 'data pipeline', which refers to a chain of data processing stages [169], as illustrated in Figure 2.13 (page 95).

In healthcare settings, Clinical Data Warehouse (CDW) technologies are designed to facilitate data integration from numerous heterogeneous systems (page 98) via the import of multiple data extracts into the CDW. [171] Thus, bringing disparate data together in a single repository, to be organised and linked to create an integrated EMR and enable data analysis. [116, 171] Hence, the Top-Down (Data-driven) approach, as per Inmon's DW Architecture [173] (page 96) was most appropriate for this research.

Additionally, the author worked with Data Warehouse Team at ICHNT to setup a CDW for this study, which ensured that data were only accessible under an appropriate level of security, technical and governance controls, deemed trustworthy by stakeholders. [169]

Furthermore, the relational database management system Microsoft SQL Server 2014, [233] was used for database management, linkage and integration. SQL is a structured query language and is the formal International Organisation for Standardisation de facto language for database management. [122, 124]

SQL facilitates communication between humans and databases [101] by permitting the querying of several relational tables within one script. For example, (Figure 2.6, page 68) also shown below, the SQL to identify all female patients with the surname 'Brown' would be written as:

```
SELECT      PatientID, Surname, Forename, DateOfBirth, Gender, Ethnicity
FROM        PatientTable
WHERE       Gender = 'Female' AND Surname = 'Brown';
```

Thus, for this study, the author created bespoke queries within each respective CIS, that included the following data elements, detailed in Table 5.5 (page 213).

- Identification numbers: Episode ID, Patient ID, NHS Number
- Quasi-identifiers: Surname, Forename, Sex, Date of birth, Postcode
- Episode dates/times: Arrival Date, Arrival Time, Depart Date, Depart Time
- Episode details: Hospital site, Clinical stream, Outcome

For the IPC dataset, the total continuous stay of a patient using a hospital bed, i.e. the 'Hospital Provider Spell' [312] with an admission method via ED, coded as '21' within the NHS data dictionary was used as selection criteria for the query. [61]

4.4.2 EMR record structure

No single medical record structure will suit every purpose, [118] and the environment in which an EMR is used will influence how/what information is recorded. Amongst the three CIS, two distinct categories of EMR were identified.

The UCC and ED databases contained EPR structures, capturing the periodic episode of care provided to a patient within one institution, i.e. treatment for a condition at a hospital. [10] EPR data are arranged in a chronological manner, and identified by date/time; thus, EPRs are data-oriented and provide little structure beyond a timestamp [118]

However, records in the IPC database were structured similar to an EHR, in which each patient is initially allocated a unique Patient ID that links all subsequent patient admissions. In turn, each 'spell' of admission is allocated/identified by a unique spell ID. Thus, it is possible to trace all admissions a patient may have had over time, via a One-to-many relationship. [10, 117]

4.5 Data quality

The previous steps facilitated the creation of a CDW, containing three distinct datasets (UCC, ED and IPC) within the time frame for this study. Prior to considering the linkage of these datasets, the quality of data must be assessed by considering "are these data fit for the purpose intended?". [184]

4.5.1 Data quality assessment

Addresses objective 5: To determine if data are fit for linkage purposes by assessing the quality of data.

Data are not just collected unintentionally; they have a purpose to exist and must be utilised in a manner which fulfils that purpose. Thus, data can be considered to 'have quality' if they satisfy the requirements of their intended use, and to achieve this data must be trusted and understood. [188]

Data Quality (DQ) encompasses several dimensions, and is considered to be the degree to which data items are accurate, complete, relevant, timely, sufficiently detailed, and appropriately represented whilst retaining sufficient contextual information to support the decision-making process. [26]

For this research, DQ dimensions: Completeness, Uniqueness, Validity, Accuracy, Consistency and Timeliness, were utilised to assess select data elements identified within the three data extracts held in the CDW. Table 2.19 (page 106), provides descriptions and calculations for each of these DQ dimensions.

Data from multiple sources, recorded by different: organisations, systems, locations, time points or data entry modes, will most likely be inconsistent. Additionally, variations in data coding across disciplines will affect its consistency between different databases. Therefore, careful mapping between different formats and encoding is required before any matching can be attempted. [177]

Data profiling is defined as the application of data analysis techniques to existing data stores to determine the actual content, structure, and quality of the data. As such, metadata is essential for determining inaccuracies, i.e. “you cannot tell if something is wrong unless you can define what being right is”. [188] Therefore, cross-referencing against data dictionaries and building a repository of information gathered from data quality facts will help create a profile for chosen data elements.

4.5.1.1 Reference table approach

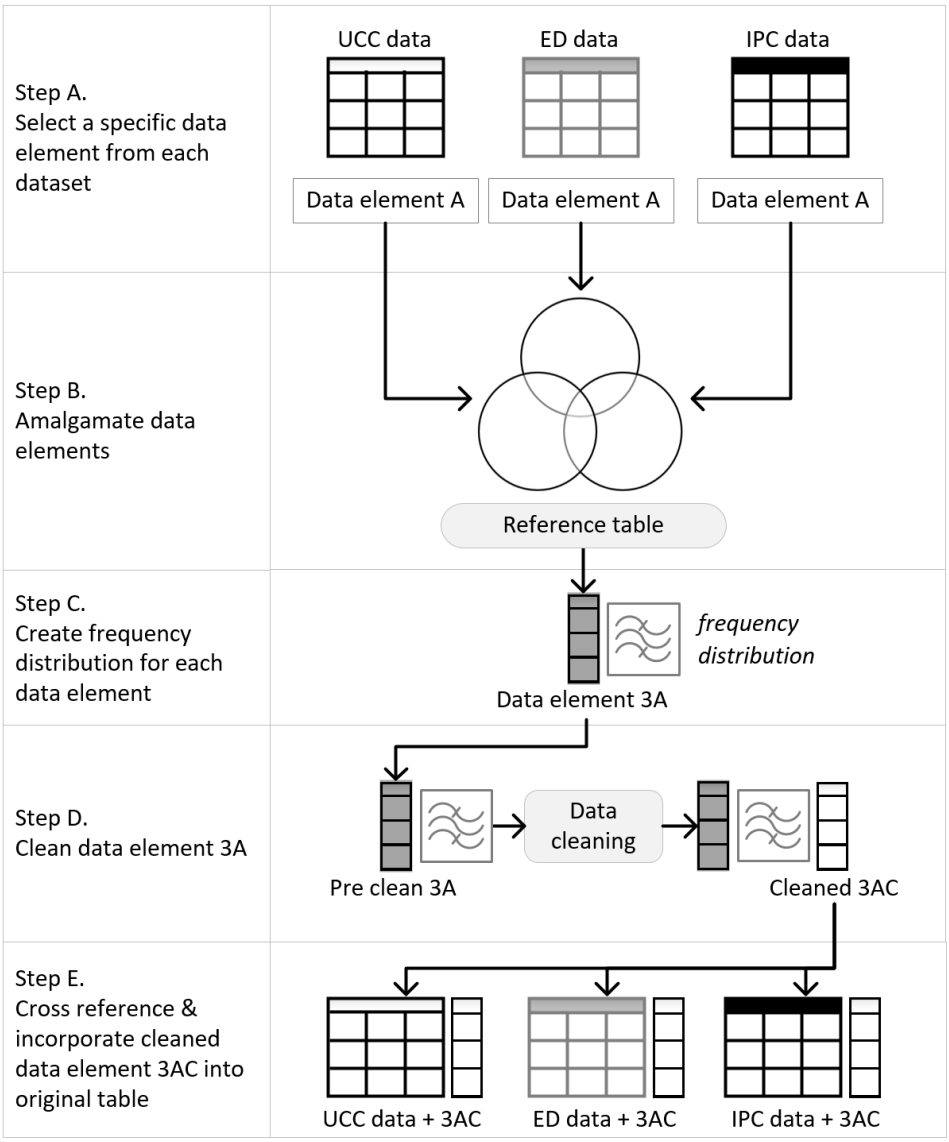


Figure 4.1 Data element cleaning via the use of reference tables

The approach outlined in Figure 4.1 (page 182) was applied to the data elements within each of the CDW datasets. For example, one specific data element is selected from all three datasets, e.g. [UCC.A], [ED.A], and [IPC.A] and amalgamated into one new reference table [MergedRefTable.3A].

Thereafter, calculating the frequency for each data element results in unique values and their corresponding frequencies, e.g. item10, frequency 2; item 11, frequency 98.

Uncommon data items have lower frequencies and can be readily identified, Step D then assesses each data item within the [MergedRefTable.3A], allocating an additional corresponding cleaned data element [MergedRefTable.3AC]. The new, cleaned data element is cross-referenced back into each original dataset (Step E). The benefit of the reference table approach is that every single data item from across all three datasets have been consolidated and cleaned in one instance, thus, maintaining consistency in approach and addressing large datasets.

4.5.1.2 Inspection of patient forename & surname

The surname is a person's family name, and it is commonly used to describe family, clan, tribal group, or marital association, [313] whereas a person's forename is the given name of that person commonly used to address them. [314]

The data elements for patient surname and forename are free text, and as such are susceptible to input errors. Hence, all keyboard characters that are numerals or non-alphanumeric (symbols) require identification and subsequent removal. [315] To ensure consistency, all names were set to lowercase and spaces removed, leaving only alphabetic characters from a to z.

Both the surname and forename were addressed separately by firstly using the reference table approach outlined in Figure 4.1 (page 182) and subsequently assessing the data quality using appropriate dimensions listed in Table 2.19 (page 106).

Furthermore, as it may be time-consuming and impractical to search for and identify every possible entry, 'wild cards' are used. In computing, a 'wildcard' is defined as a special character that will match any character or sequence of characters in a search. [316]

Depending on the programming language being used, wildcards tend to be either an asterisk (*) or percentage symbol (%). In SQL script, the percentage symbol is used, e.g. when searching for the term 'The%' will provide results such as 'The, Them, Then, There, Therefore' etc. Hence, all text possibilities beginning with 'The' will be identified.

Additionally, searches must consider, numerals intermingled with the name: e.g. 0'Surname (zero'Surname); Date of birth incorrectly recorded; non-alphanumeric symbols: e.g. ".,';[]/\()+~`!@#$%^&*_<>?"; removal of all leading, intermittent and trailing spaces; and the presence of 'Test' cases. [317]

Criteria to address	Example search terms
Unknown & derivatives	Name Like 'Unk%' OR 'Ukn%' OR 'Anony%'
Not known OR Not given	Name Like 'Not%'
Titles	Name In ('Mr','Mrs','Miss','Ms', 'Dr','Prof ', 'Baby')
Test entries	Name Like 'Test%' OR 'Dummy%' OR 'Dustbin' OR 'Patient%'
Departments	Name In ('aehh', 'aecx', 'ae', 'symphony')
Numbers	Name Like '%[0-9]%'
Special characters OR Spaces	Name Like '%[^a-zA-Z0-9]%'

4.5.1.3 Inspection of patient date of birth

The date of birth of the patient is usually stored in the format DD-MM-CCYY. [318] Firstly, the reference table approach outlined in Figure 4.1 (page 182) was employed, followed by an assessment of data quality using appropriate dimensions, as listed in Table 2.19 (page 106). The NHS data dictionary states that the patient date of birth must range between 01/01/1880 to 31/12/2999 in order to pass validation. [319]

4.5.1.4 Inspection of patient gender

The sex/gender of a person is identified as either Male or Female. [320] However, it may be coded as 1 = Male, 2 = Female, or abbreviated to M or F, or written in full as Male or Female. Additionally, there may be other options specific to each CIS, e.g. Unknown, or Not recorded. Referring to the data dictionaries for each dataset revealed that the data element sex, had been stored as a string; thus the reference table approach outlined in Figure 4.1 (page 182) and subsequent assessment of data quality using appropriate dimensions listed in Table 2.19 (page 106) were applied.

4.5.1.5 Inspection of the patient postcode

The data element postcode of the residential address is frequently recorded at the initial registration stage, along with other demographic details. [321] Postcodes are alphanumeric in structure, and are therefore stored as strings. Hence, it was possible to use the reference table approach outlined in Figure 4.1 (page 182) and assess data quality using appropriate dimensions, as listed in Table 2.19 (page 106). In addition, data item searches must detect inconsistent entries, e.g. unknowns, special

characters or just ‘numbers’, as in non-UK overseas codes. Within the postcode reference table, all intermittent spaces were removed, and the text case set to UPPERCASE and an updated data element ‘Pc1’ was created.

Criteria to address	Example search terms
Unknown	Postcode Like ‘Unk%’
No fixed address	Postcode Like ‘NF%’
Just numbers	Postcode Like ‘[0-9]%’
Special characters	Postcode Like '%[^a-zA-Z0-9]%'

Furthermore, NHS England recommends that pseudo postcodes should be coded as follows: ‘No Fixed Abode’ (ZZ99 3VZ); ‘Not Known’ (ZZ99 3WZ); ‘UK not specified’ (Z99 3CZ). [322, 323] Thus, to provide additional context to the postcodes, appropriate reference data tables were obtained from the Office for National Statistics [324] NHS Postcode Directory (2016 Edition) and mapped to the cleaned data element ‘Pc1’.

Additionally, for all UK postcodes, the Lower Layer Super Output Area name (LSOA) were obtained and linked to the data element ‘Pc1’. [325] This additional information helped to differentiate between UK, Overseas and Unmapped postcodes. For example, within the dataset, a postcode of ZZ994MZ maps to the country Austria (Europe) within the NHS Postcode Directory.

4.5.1.6 Inspection of Episode Arrival/Depart date/time

The arrival date is the date the patient self-presented or arrived via an ambulance to the ED. [326] Similarly, the arrival time is when the patient self-presented or arrived in an ambulance at ED, recorded using the 24-hour clock format. [327]

Additionally, the depart date is the date that a patient leaves an ED after the attendance has concluded, [328] and the depart time (recorded using the 24-hour clock format) is the time that a patient leaves ED after the ED attendance has concluded. [329]

Within each dataset, the ‘Arrival Date/Time’ and ‘Depart Date/Time’ data elements signify the exact time when a patient episode of care commenced and concluded. However, referring to the data dictionary for each dataset, revealed that within ED and IPC databases, both the arrival and depart data elements of date and time were being stored as two separate data elements, whereas UCC database stored date/time as one combined data element.

For example, if a patient arrived in ED at 15/05/2015 23:45 and left the department at 16/05/2015 00:30, the actual time in department 45 minutes. However, calculating the difference between just the dates would result in a 24-hour difference. By combining the date and time data elements, it is possible to calculate the exact difference between the arrival and departure time for each patient episode. Therefore, a new data element was created using the following SQL script to combine the 'Date' with the 'Time' for both the ED and IPC datasets:

```
ED:CONVERT (datetime, [TableED].[ArrDate], 103) + CONVERT (nvarchar, [TableED].[ArrTime], 103)
IPC:CONVERT (datetime, [TableAU].[ArrDate], 103) + ' ' + CONVERT (char(8), dbo.T13Ic.ArrT0, 8)
```

The code differs slightly due to the time data element structure, within the respective CIS. Within the code, '103' is the British Standard for Date format (DD/MM/YYYY) and '8' refers to time format (hh:mm:ss). [330]

All three datasets were updated to ensure the arrival and depart data elements possessed the same structure; this enabled an assessment of each element separately, via application of the reference table approach outlined in Figure 4.1 (page 182) and further assessment of data quality using appropriate dimensions listed in Table 2.19 (page 106).

4.6 Data linkage

Addresses objective 6: To differentiate the effect of deterministic and non-deterministic data linkage strategies with respect to the numbers of correctly matched records.

To differentiate empirically whether the selection and application of a particular linkage strategy influenced the number of correctly identified records belonging to the same individual across multiple datasets, the hypothesis (page 21) was tested using the Two Proportion Z-Test, which permits the comparison of proportions from two independent samples to see if they are the same. [179]

H_0: *There is no difference between using deterministic or probabilistic record linkage techniques.*
H_1: *There is a difference between using deterministic or probabilistic record linkage techniques.*

This study will apply both strategies to the exact same datasets and compare the potential of each strategy at correctly identifying a patient across multiple datasets in order to facilitate patient deduplication. Furthermore, to gain an advanced understanding of record linkage strategies, the author attended appropriate courses to support this research. [331]

4.6.1 Deterministic linkage strategy

In order to achieve DRL, the HES 'Step-based DRL' strategy [206] using the NHS Number and patient quasi-identifiers date of birth, gender and postcode (detailed on page 115) was applied to the datasets. This linkage strategy, as recognised by NHS England, has proven applicability in identifying potential matches between two datasets.

4.6.2 Hybrid (DRL & PRL) linkage strategy

The literature identified Hybrid strategies [201, 220, 244] which use a combination of DRL and PRL linkage strategies using quasi-identifiers: surname, forename, gender, date of birth and postcode. Additionally, the PRL approach employs Blocking, Soundex, Combination ID, Splitting the date of birth, String comparator, and Relative frequency, all of which are detailed on page 121 of this thesis.

For patients receiving treatment via multiple departments, the overall longitudinal record of care of the patient's journey was fragmented and held within silos by each department. The activity diagram, Figure 5.1 (page 203) identified that patients may present to the UCC, be referred to the ED, with some patients requiring admission to IPC. Therefore, to identify those patients that were seen throughout more than one service; data extracts were obtained from the CIS of each service, and will require linkage in order to create a longitudinal patient record. [114]

4.6.2.1 Creation of Quasi-identifier

In the absence of a global unique identification number, a patient may be identified using only quasi-identifiers, i.e. patient demographic information, forename, surname, sex, date of birth and postcode. [202, 217, 220, 221] The literature review section 2.6 (page 112) identified a technique whereby data elements could be combined to create a new combination identifier.

Additionally, the NHS Strategic Tracing Service (NSTS) report that in the absence of the NHS number, a minimum trace would require Surname, Forename, Date of Birth and the first part of postcode. [332] As these data elements are available within all three datasets, a Quasi-identification number (QID) was created by concatenation of Surname, Forename, Date of birth and Sex. This new data element QID was incorporated within each dataset, and described within Figure 4.2 (page 188), thus, establishing the ability to uniquely identify patients across all three datasets.

$$\boldsymbol{Quasi\ ID = Concatenate\ (Surname + Forename + Date\ of\ Birth + Sex)}$$

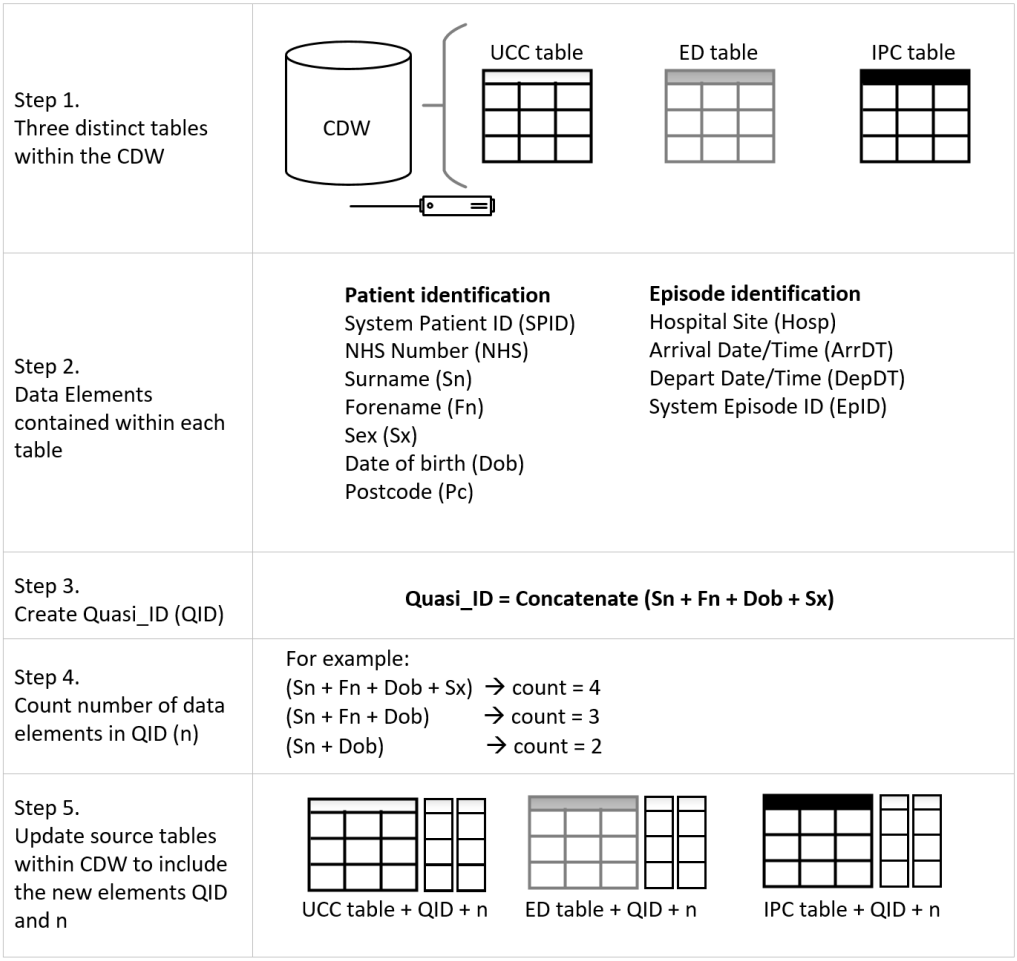


Figure 4.2 Incorporation of Quasi-identifier.

4.6.2.2 Creation of reference tables for multiple IDs

Figure 4.3 (page 189) outlines the process of creating reference tables for each ID data element, for example: selecting the QID from all three tables and amalgamating them to create a new reference table, in which each QID is listed uniquely with its corresponding frequency. This value represents the total number of times the QID appears across all three datasets.

Additionally, a binary 'source table' indicator is allocated as per Step 8a Example, and a count which indicates the number of tables within which the unique QID appears, this information will be further utilised to facilitate deduplication processes.

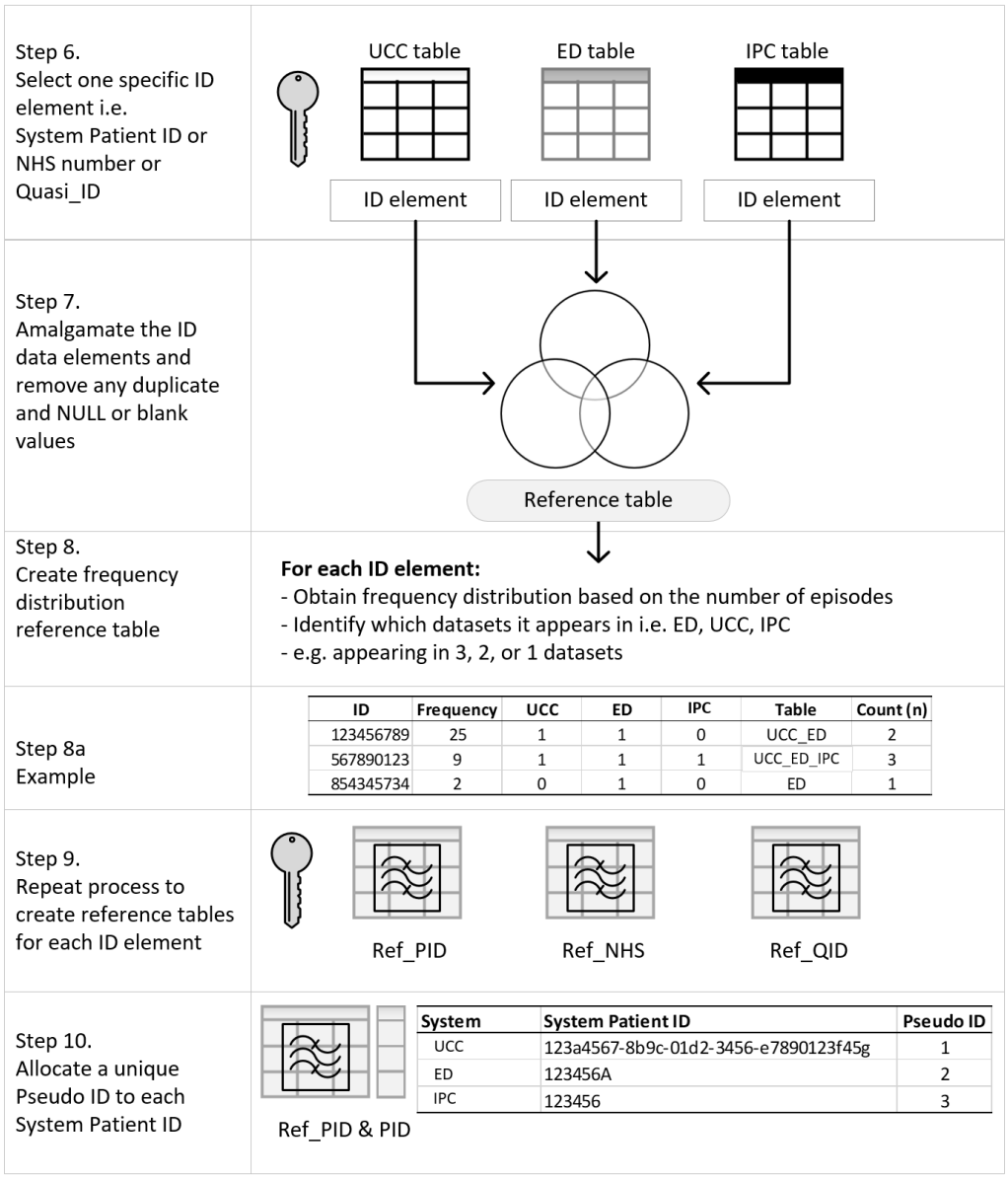


Figure 4.3 Reference tables for each ID type data element.

Additionally, in Step 10, each System-specific patient ID is heterogeneous and is only applicable to its source, i.e. UCC or ED or IPC. Therefore, a corresponding Pseudo ID, comprised of numerical incremental values, was allocated to each record and stored within the Ref_PID table. This process enabled the Pseudo ID to act as a universal identifier across all three disparate datasets.

4.6.2.3 Data element variations per unique ID

Figure 4.3 (page 189) resulted in the creation of three distinct reference tables based on Ref_PID (Pseudo ID), Ref_NHS (NHS number) and Ref_QID (Quasi-ID), and enabled three potential approaches to identify a unique individual across all three datasets. However, two of three datasets hold seven years of data, each dataset contains thousands of records, and individual patients may

have multiple attendances and EPRs with dissimilar details, which may/may not include transcription errors, thereby increasing the difficulty in ascertaining the 'correct' details for a patient. For example, when looking at actual values recorded for the patient surname, it is possible to have more than one value, perhaps due to interpretation or typographic error or patient preference.

Thus, values for surname could be recorded as Smith, Smyth or even Smithe, and may contain even more potential variations or variants. Figure 4.4 (page 190) illustrates this point, whereby a patient may have three possible surname variants Smith, Smyth or Smithe, and the difficulty is knowing which variant to use. Although the most commonly used surname identified as 'Smith' from US Census (2010) and Scottish Records, [333, 334] it is not possible to find such details for all names, especially ethnic minority names.

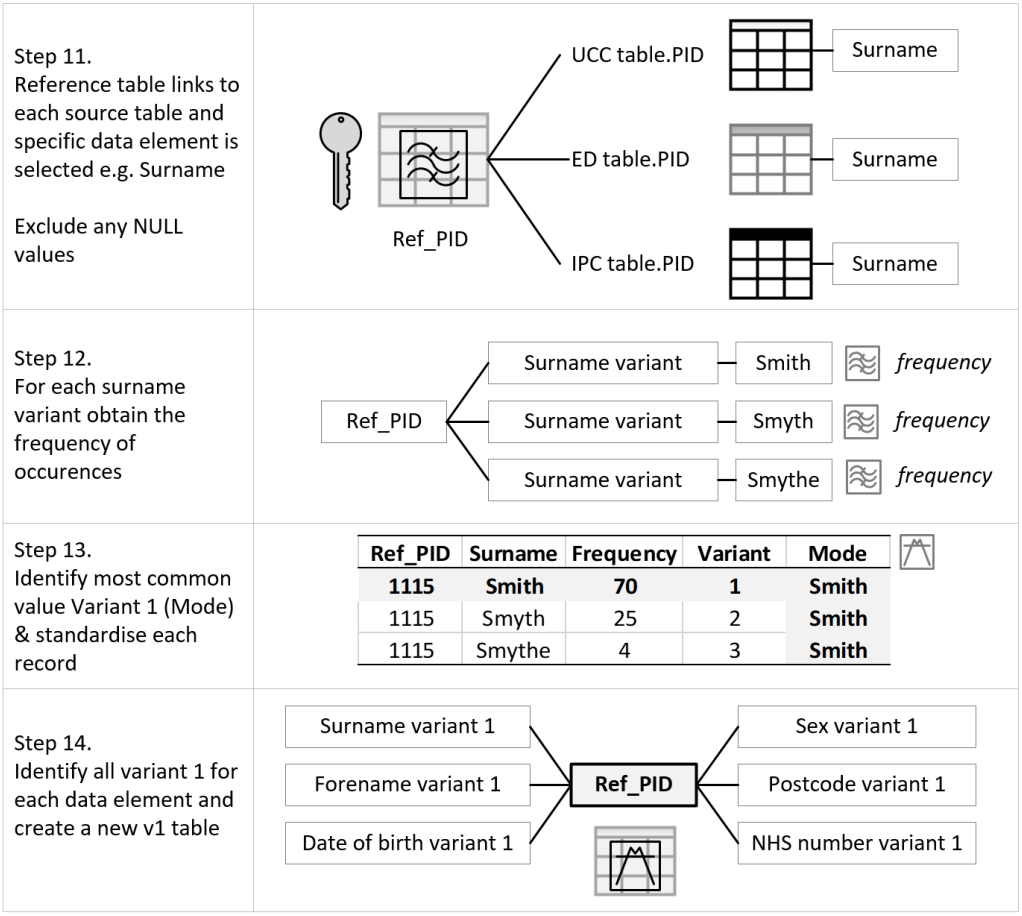


Figure 4.4 Example of identifying the number of Surname variants per unique ID

However, one approach to overcoming this issue is to calculate the frequency for each specific surname variant, to reveal the modal value. The mode reveals the name most commonly used by the patient, in this example variant 1, i.e. 'Smith', and holds superiority over the other two name

variants. Therefore, the modal value was selected as the principal value and allocated to each record in a hierarchal manner, to provide a means of homogenising the element surname, whilst retaining the original surname variants for future validation purposes.

Thus, deduplication via the mode (based on variant frequency) permits the selection of the most predominant value for each data element, i.e. the most commonly used value prevails, and uncommon values or errors with low frequencies become redundant, thereby consolidating all variants and duplicates per each ID.

Additionally, by repeating this process for each of the three IDs and their associated data elements, (Step 14), it is possible to identify and link all the Variant 1 (Mode) entries to create a new reference table that contains one record per ID, devoid of duplication.

Furthermore, individual patients were triangulated [335] within three disparate patient identifiers that resided within each of three datasets, i.e. Pseudo ID, NHS number, and Quasi-ID. Figure 4.5 (page 191) illustrates how a unique patient may be identified by using any of the three IDs, whilst additionally highlighting the interrelationships between each of these three IDs, further complicating the process of selecting the most appropriate ID for linkage purposes. For example, one patient may have two NHS numbers, three PIDs and one QID, resulting in duplication.

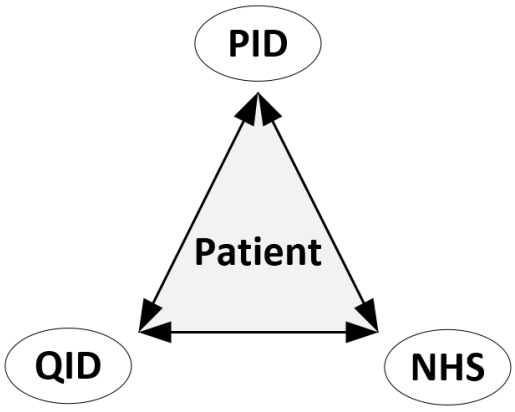


Figure 4.5 Triangulation of a patient within distinct IDs.

Nonetheless, to facilitate record linkage, one must ensure that there are no duplicate values present within the tables to be linked and that each ID appears only once, i.e. a ratio of 1:1:1 for each identifier PID:NHS:QID.

Therefore, due to the presence of missing values and low levels of completeness, the NHS number could not be used as the overarching ID, whilst the Pseudo ID contained high levels of variability making it unsuitable for linkage purposes. However, as the Quasi-ID was a concatenation of Surname, Forename, Date of Birth and Sex, it was able to distinguish and discriminate between individual patients accurately and was therefore selected as the overarching ID across all three datasets to be used within the subsequent record linkage process. The results for deduplication via ID have been presented in Figure 5.6 (page 226).

4.6.2.4 Hybrid PRL

The methods thus far, have employed DRL strategies that have resulted in the creation of one amalgamated dataset that does not contain any duplicate records, whilst additionally facilitating the selection of the unique Quasi-ID as the predominant ID to be utilised for subsequent PRL purposes.

An overview of the Fellegi & Sunter approach (page 112), along with corresponding linkage strategies to facilitate and improve the accuracy of PRL were identified within the literature. These strategies include PRL, Blocking, Soundex, Combination ID (used to create the Quasi-ID), Splitting the date of birth, Jaro-Winkler string comparator, and Relative frequency, all of which are detailed on page 121 of this thesis.

The literature identified blocking techniques utilising the Soundex algorithm to overcome typographical errors; hence, this algorithm was applied to Surname and Forename, respectively. The Soundex algorithm is available as a pre-programmed function within the Microsoft SQL Server application. Furthermore, the date of birth was split into component parts, day, month and year and stored as separate data elements ([DD], [MM], [YYYY]) to account for any typographical errors and artificially increase the number of data elements available when comparing records.

The data elements surname and forename contained values that showed a high degree of variation due to the presence of ethnic minority names. As uncommon names possess more distinguishing power, applying relative frequencies would adjust the match weight upwards and serve to enhance the performance of PRL and improve specificity. [213]

Additionally, when comparing names, e.g. Grey = Gray, the Fellegi-Sunter algorithm results in a binary outcome that would result in a zero match. However, using the Jaro-Winkler algorithm on the same comparison would result in a value of 0.94, as it accounted for the number of similarities and

transpositions. The Jaro-Winkler algorithm and source code were freely available as part of the SimMetrics library [336], which integrated into Microsoft SQL Server as a Common Language Runtime function, enabling its application to the dataset.

Nonetheless, a unique patient could potentially have between zero and 12 postcodes, which further introduced a high degree of variability and potential for duplication of records. For example, a person with eight associated postcodes would have eight database records with exactly the same quasi-identifying data elements (e.g. surname, forename) being repeated, with the exception of the postcode. This would result in the presence of duplicate entries, artificially increasing the number of comparisons for clerical review, whilst requiring more computer-processing time. However, the transposition (reshaping from long to wide) of postcodes for each patient record would eliminate all duplication of quasi-identifying values, and serve to reduce the number of record comparisons and processing time.

For this study, prior to undertaking PRL, the Match (m) and Unmatch (u) probabilities were calculated by a comparison of classification pairs, matching the Quasi-ID with each specific data element (illustrated on the right); actual probabilities listed in Table 4.1 (page 193).

$$\{QID\} \times \left\{ \begin{array}{c} (Surname), \\ (Forename), \\ (Sex), \\ (BirthYear), \\ (BirthMonth), \\ (BirthDay), \\ (Postcode), \\ (Soundex{:}\,Surname), \\ (Soundex{:}\,Forename) \end{array} \right\}$$

Table 4.1 M and U probabilities used within the PRL approach (Step 19)

Probability	$Quasi\ ID \times$	Alias	Soundex Surname	Soundex Forename
Match (m)	*Surname*	mpSn	0.999999999	0.999776475084182
	Forename	mpFn	0.999935095373742	0.999999999
	Sex	mpSx	0.999471147489754	0.999471231381937
	BirthYear	mpBy	0.9980696883376	0.998069994544069
	BirthMonth	mpBm	0.9980696883376	0.998069994544069
	BirthDay	mpBd	0.9980696883376	0.998069994544069
	Postcode	mpPc	0.967112585487806	0.967014452207729
	SndxForename	mpFnx	0.999935095373742	-- exclude
	SndxSurname	mpSnx	-- exclude	0.999776475084182
Unmatch (u)	*Surname*	upSn	(Relative frequency table)	(Relative frequency table)
	Forename	upFn	(Relative frequency table)	(Relative frequency table)
	Sex	upSx	0.501009403107583	0.71915198661532
	BirthYear	upBy	0.0176661715012116	0.0198773432312198
	BirthMonth	upBm	0.0839123198076517	0.0835234575842072
	BirthDay	upBd	0.0343990824284157	0.0329790178956482
	Postcode	upPc	0.000904437415691017	0.000232565441090737
	SndxForename	upFnx	0.00582613192392134	-- exclude
	SndxSurname	upSnx	-- exclude	0.0017628078309313

Importantly, as the m and u probabilities are not available from any sources of literature, they must be calculated empirically via a comparison of classification pairs from the study dataset, [198, 199] as described in Table 2.23, page 119 of this thesis.

Step 15. Amalgamated Table A with Quasi ID and all unique records

Table A

Data elements

						Birth			Postcodes		
QID	Surname	Soundex	Forename	Soundex	Sex	Year	Month	Day	1	2	3
5	Sample	S514	Patient	P353	M	1999	5	2	Pc1	Pc2	Pc3

Step 16. Create u-probability distribution for Surnames and Forename



Step 17. Create clone Table B and join on Soundex Surname



Step 18. Calculate match status binary comparison vector (match = 1, else 0)

	a	b	c	d	e	f	g	h	i	j	k	l
Table	QID	Surname	Soundex	Forename	Soundex	Sex	Year	Month	Day	1	2	3
A	5	Sample	S514	Patient	P353	M	1999	5	2	Pc1	Pc2	Pc3
B	5	Sample	S514	Patient	P353	M	1999	2	5	Pc1	Pc2	Pc3
Vectors	1	1	1	1	1	1	1	0	0	1	1	1

Step 19. Calculate probabilities for all data elements, except blocking data element

Match QID

Match Data Element	1	0	∑
1	a	b	a+b
0	c	d	c+d
∑	a+c	b+d	n

Probability	Alias	Calculation
Match (m)	Sensitivity	a/(a+c)
Unmatch (u)	1 - Specificity	1 -[d/(b+d)]

Step 20. Calculate match weights (log likelihood ratio) from result of vector (in Step 18)

If vector = 1 $w = log_2\left[\frac{m}{u}\right]$

If vector = 0 $w = log_2\left[\frac{(1-m)}{(1-u)}\right]$

Step 21. Account for complex agreement patterns in Surname & Forename

String comparator for Surname and Forename
Use Jaro Winkler algorithm to compare names, to take into account spelling differences e.g Grey and Gray = 0.87

Step 22. Calculate total weights

∑ weight(all data fields)

Step 23. Calculate Probability

exp (∑ weight)

Step 24. Create PRL table

	a	b	c	d	e	f	g	h	i	j	k	l
Table	QID	Surname	Soundex	Forename	Soundex	Sex	Year	Month	Day	1	2	3
A	5	Sample	S514	Patient	P353	M	1999	5	2	Pc1	Pc2	Pc3
B	5	Sample	S514	Patient	P353	M	1999	2	5	Pc1	Pc2	Pc3
Vectors	1	1	1	1	1	1	1	0	0	1	1	1
m Probability	m	m	m	m	m	m	m	m	m	m	m	m
u Probability	u	u	u	u	u	u	u	u	u	u	u	u
u Name Table		u Surname		u Forename								
log weight	wt(m)	wt(m)	wt(m)	wt(m)	wt(m)	wt(m)	wt(m)	wt(u)	wt(u)	wt(m)	wt(m)	wt(m)
JW algorithm		JW (SnA,SnB)		JW (FnA,FnB)								
Final weights	wt.a	jw(wt.b)	wt.c	jw(wt.d)	wt.e	wt.f	wt.g	wt.h	wt.i	wt.j	wt.k	wt.l
Total weight	∑ wt											
Probability	exp (∑ wt)											

Note: Weights for the blocking data element are not included in the total weight

Figure 4.6 Application of probabilistic record linkage

The PRL stage of the Hybrid algorithm, as illustrated within Figure 4.6 (page 194) was implemented using the amalgamated dataset using two distinct blocking strategies, strategy one using Soundex Surname and strategy two using Soundex Forename. In each case, it was important to ensure that each of these blocking data elements did not contain any missing (NULL) values and that their results were not included in calculations of total match weight. For example, if blocking using Soundex Surname, then do not include the m or u probabilities for Soundex Surname in the PRL process.

Within strategy one, the dataset was cross-matched with a virtual clone of itself to create a Cartesian product that joined the tables using Soundex Surname as the blocking strategy. This resulted in the matching of every single record to create matched pairs and corresponding match weights and probabilities and further permitted classification of these matches into three categories Matched, Unmatched or Clerical Review. Whereby, Matched records incorporated those comparison pairs in which the QID and additional details matched, in contrast, Unmatched records included those records where neither the QID nor details matched.

However, Clerical Review encompassed the missed or potential matches, i.e. those record pairs in which although the QID did not match, other details matched to an extent, possessing a relatively high total weight and associated probability. Therefore, to be conservative, only record pairs where the QID did not match, but the match probability was greater than or equal to 0.8 were included for Clerical Review. Additionally, strategy two involved repeating the process described for strategy one, but instead joining the tables using Soundex Forename as the blocking strategy.

Moreover, strategy one and two had identified potential record pairs that were appropriate for Clerical Review, and to simplify the review process, results from each step were combined and all duplicates identified and removed, e.g. $Strategy\_One(Z \times A) \equiv Strategy\_Two(A \times Z)$.

Thus, the Hybrid approach enabled data linkage and deduplication via the initial use of DRL techniques followed by the application of PRL techniques to identify additional missed matches. However, it should be noted that due to the large numbers of records, the computation time for the PRL process may be considerable, i.e. between one to 12 hours. Therefore, preliminary coding and testing on a small subset of data is highly recommended.

4.6.2.5 Master Patient Table based on Quasi-ID

The application of the Hybrid linkage strategy resulted in the creation of one Master Patient Table containing unique patient records and an associated Master Patient Identifier based on the Quasi-ID, which readily identified any patient from across all three UCC, ED and IPC datasets. Data within the Master table had undergone cleaning, standardising, and assessment processes to ensure it did not contain any duplicate records. The Master Patient Table was utilised as a 'Junction' table to link the three disparate datasets via the Quasi-ID. Thus, creating a Linked dataset and establishing a longitudinal record of care for patients that were seen throughout all three departments.

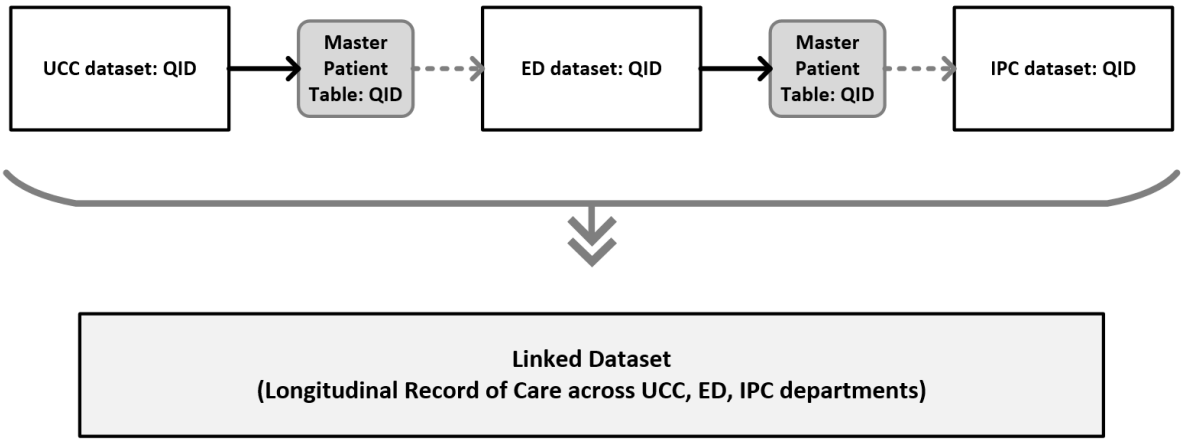


Figure 4.7 Creation of Linked dataset via linkage of UCC, ED and IPC datasets with Master Patient Table.

4.7 Linked dataset analysis

The previous step resulted in the creation of a Linked longitudinal dataset, stored within the CDW, spanning the entire patient pathway from the UCC, through ED to IPC services. Furthermore, to demonstrate practical application of the linked dataset, analyses which would not have been possible using unlinked datasets, are presented henceforth.

4.7.1 Transition times

Addresses objective 7: To calculate the transition time for patient transfer between urgent and emergency care services

In certain cases, walk-in patients register at the UCC and are initially seen by the GP streamer, but if their condition warrants treatment in ED, they are provided with a referral letter. Such patients have to queue again at reception to register for ED services. The additional waiting time required for a

patient, from the end of one episode of care within the UCC database, to commencing a new episode of care within the ED database, is defined as the 'Transition time' (minutes) within the context of this study, which is the time in between departmental patient transfers that cannot be calculated from within any one database system.

However, it contributes to the total amount of time a patient spends in the department and is affected by busy periods. It can be calculated using standard SQL functions such as Date Difference, where the data element is in the YYYY/MM/DD HH:MM:SS format, for example in SQL the function DATEDIFF calculates the difference in seconds between the UCC episode End date/time and the ED episode Start date/time:

```
DATEDIFF(second,[UCC.EpisodeEndDT],[ED.EpisodeStartDT])
```

Additionally, comparisons of the mean time (minutes) between CXH and HH sites were undertaken:

- Transition Time from UCC to ED
- Referral time from ED to IPC
- Referral time from ED to UCC

The underlying assumption is that the two groups come from populations with the same mean, i.e. there will be no significant difference between groups. Thus, the t-test for two independent groups was conducted to determine whether there was a statistically significant difference between the mean transition time between the two groups, CXH and HH. [179]

4.7.2 ED four-hour standard

Addresses objective 8: To distinguish the effect of utilising Linked and Unlinked datasets upon the ED 4 hour standard.

The four-hour standard is a Clinical Quality Indicator [68] focusing on the total time a patient spends within ED (minutes) from Arrival time to Depart time, as detailed on page 34 of this thesis. The statistics reported for this indicator comprise:

- The Median, i.e. 50th percentile
- The 95th percentile
- The single longest total time spent by patients in ED

However, when calculating the total time in ED, in some episodes chronological discrepancies between the date/time as recorded within different datasets were identified, e.g. for patients referred from ED to IPC, the ED episode End date/time was greater than the IPC episode Start date/time.

Therefore, within the CDW Linked dataset, this was reconciled using a SQL conditional statement that automatically selected the Minimum date/time value when calculating the total time in ED:

```
DATEDIFF(second,[ED.EpisodeStartDT],
[CASE  WHEN [ED.EpisodeEndDT]>[IPC.EpisodeStartDT]
       THEN [IPC.EpisodeStartDT]
       ELSE [ED.EpisodeEndDT] END])
```

Moreover, the objective is to calculate the aforementioned statistics for this indicator, using two distinct ED datasets, the CDW Linked dataset and an ED Unlinked dataset, both with exactly the same selection criteria. The null hypothesis for this objective states there will be no difference in the results for the total time in department (minutes) between the two datasets.

In order to test this hypothesis, the results will be compared using the t-test for two independent groups [179] to determine if there is a statistically significant difference between the utilisation of the Linked dataset or Unlinked dataset and the results will be presented separately by hospital site.

4.7.3 Comparison of admission rate pre/post UCC

Addresses objective 9: To compare the rate of emergency patient admissions pre/post UCC implementation

"How do we know if change is an improvement?" [337] To answer this question, this section presents an application of the CDW Linked dataset by looking at the rate of emergency patient admissions pre/post UCC implementation.

To facilitate this analysis, specific data elements were selected from the CDW linked dataset, which included patient age, sex, postcode; episode arrival date, and outcome, indicating whether a patient was admitted or discharged. The data element age was calculated from the difference (in years) between birth date and arrival date and grouped into the following categories: 0 to 14, 15 to 24, 25 to 39, 40 - 59, 60 - 79 and 80+ years.

Furthermore, for each patient episode, the postcode data element was cross-referenced against the Index of Multiple Deprivation (IMD) scores, to obtain the appropriate socio-economic status quintiles. [297]

Additionally, the duration of admission was calculated (hours) from Admission Start date/time to Admission End date/time and further categorised as Short stay, i.e. less than two days (< 48 hours), and Long stay, i.e. greater than or equal to two days (≥ 48 hours). [78]

Due to differences in hospital profiles, each hospital site CXH and HH was analysed separately, hence one dataset created per each site. Each dataset was aggregated by calendar year and month, patient sex and age band. The study period, i.e. 01st January 2007 to 31st December 2013, was approximately 2.5 years pre and 4.5 years post establishment of the UCCs. [38] Utilising the linked dataset overcome the issue of double-counting patients that were seen within the UCC and ED, and such patients were only counted once in the analysis.

The outcome variables were total ED attendances, where post-UCC the ED attendances included the combined attendance counts for the ED and UCC, and short-stay (< two days) and long-stay (≥ two days) emergency hospital admissions. The unit of analysis was total attendance/admission by month, age, gender and socioeconomic status.

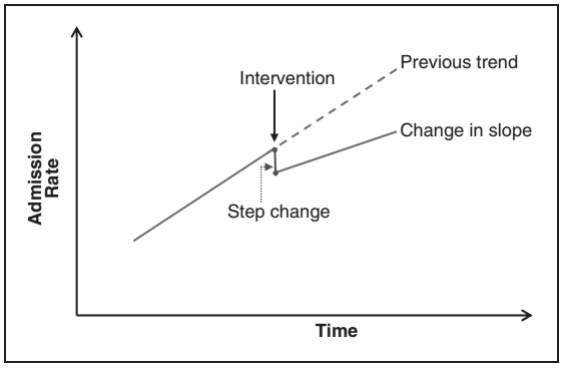


Figure 4.8 Interrupted time series design and regression analysis (Gnani et al., 2013).

An interrupted times series (ITS) study design was used to assess the effect of the UCCs on ED attendances and emergency hospital admissions. [338-340] The ITS model examines data at multiple time points over seven years and estimates the effect size of the UCCs by taking into account underlying time trends, seasonal variation and patient's characteristics. Covariates within in the model include Monthly effect (split into 12 months), Age groups (split into six groups), Gender (split

into two groups), Socioeconomic status index (split into five groups). To construct the ITS model, a dummy variable coded '0' for pre-UCC and '1' for post-UCC was used. The SQL script to identify the time frame pre/post UCC service launch:

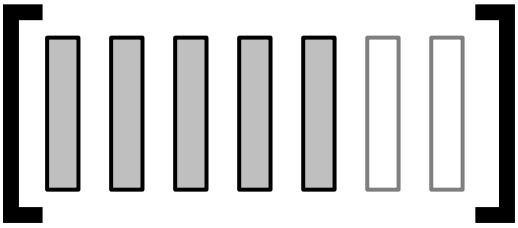
```
(CASE  WHEN "Site" = 'HH'  AND Arrival_Date >= '2009/04/23' THEN 'Post_UCC'
       WHEN "Site" = 'CXH' AND Arrival_Date >= '2009/09/07' THEN 'Post_UCC'
       ELSE 'Pre_UCC' END)
```

The ITS model included an incremental time variable, e.g. assigning '1' for the first month, '2' for the second and so on, to capture any changes in long-term secular trends. The number '0' was assigned to the month when the UCCs were implemented, and the number '1' assigned to the following month, increasing this value cumulatively for subsequent months.

The Poisson distribution is a discrete distribution, widely used in statistics for count data to describe the distribution of counts of specific events that happen randomly in time. As the outcome variable for this study are counts, a Poisson linear regression model was fitted [179], to examine the effect of UCCs on ED attendances, with the number of ED attendances as an offset variable, to examine emergency hospital admissions. The model produces changes in incidence rate ratios (IRRs) in relation to the rate projected by the underlying trend.

Multi-collinearity was tested and found that the variance inflation factor for all predictor variables was less than five, indicating reasonable independence. The non-linearity of the time trend was also tested by including a quadratic term for the time variable, which produced non-significant results, suggesting that a linear model gave an appropriate fit to the data. All statistical analyses were conducted using Stata 14 (StataCorp). [235]

Chapter 5 Results



5.1 Introduction

This chapter presents the results for this study, aligned to meet the specific objectives of this research and reflect the sequential order of processes within the conceptual framework. Thus, results for each stage, Domain understanding, Semantic interoperability, Syntactic interoperability, Data quality, Data linkage and Linked data analysis, are presented henceforth.

5.2 Domain understanding

Addresses objective 1: To explore the formation of data silos within healthcare settings.

The UML activity diagram, Figure 5.1 (page 203) visualises the patient pathway through urgent and emergency care services, with both HH and CXH sites using the same model. It focuses on the registration process, highlighting the various CISs encountered, patient identification numbers, creation of new episodes of attendance or admission, and transfer of patients between services.

The solid black circle represents an initial node when an activity is begun, whereas the final node when an activity has ended is depicted as two concentric circles that appear as a solid black circle encased by a white outline, similar to a ‘target’ or ‘bull's eye’. The arrows indicate the flow of various processes, the diamond shape represents a decision node, and the fork/join nodes depict multiple processes running concurrently. [310, 341]

Figure 5.1 (page 203) identifies four distinct lanes; three include the departments UCC, ED, IPC and one lane for Diagnostic tests incorporating the Radiology and Pathology databases, which are not covered within this research.

5.2.1 Activity diagram

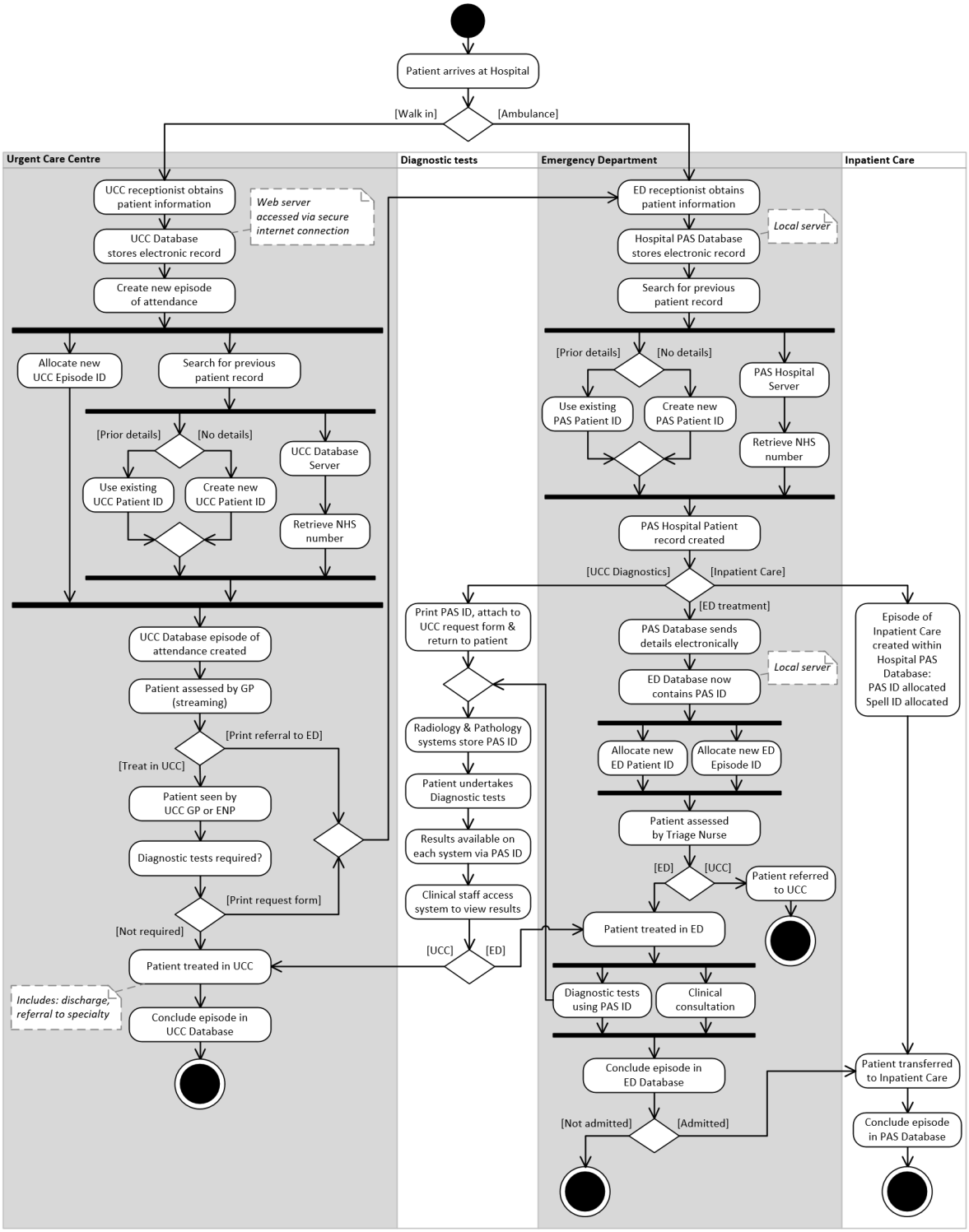


Figure 5.1 UML Activity diagram of information capture via urgent & emergency care patient pathway.

Table 5.1 Summary of information related factors identified along the urgent & emergency care pathway (Figure 5.1)

Factor	Description
Information fragments	• Each service, UCC, ED, IPC contained a bespoke database to manage the EPR for a patient episode of care. • A typical patient would encounter these databases at different stages along their treatment pathway. • Each database captured a specific piece/fragment of information about the entire patient episode of care.
Information transfer	• The UCC database was completely isolated from the ED and IPC database. • The IPC database was able to transmit an electronic message (one-way) containing patient demographic details to the ED database.
Identification numbers	• Each database was found to contain three types of identification number, two of which related to the patient, one relating to the episode of care. • The underlying database model determined the priority and uniqueness of each identification number. **Episode ID precedence** • For the UCC and ED databases, within each EMR, the 'Episode ID' held superiority, as this was the unique record identifier automatically created and allocated by the system. Secondary to this, a Patient ID, automatically created and allocated by the system, was stored in order to identify the patient. • Hence, if the same patient presented twice in a week, two completely new and different Episodes IDs would be allocated, which meant there was no way to link these episodes. • Consequently, demographic details were being re-entered for each episode, increasing the chance for transcription errors. **Patient ID precedence** • In contrast, the IPC PAS database prioritised the 'Patient ID', which held superiority and was the unique identifier for a patient. Subsequent spells of admission were allocated a 'Spell ID'. • Thus, one Patient ID was associated with many Spell IDs, which facilitated the identification and linkage of records. **The NHS number** • The UCC database provider had enabled an NHS number-checking feature. • The IPC database provider had enabled an NHS number-checking feature. • The ED database provider did not have an NHS number-checking feature, and would only populate if the NHS number was included as part of the IPC electronic message. **Additional complications** • Patients attending ED were initially allocated a temporary PAS Patient ID that was stored within the ED database. However, the PAS Database had a dedicated team who retrospectively updated the PAS database, allocating a new permanent PAS ID to each patient. This update was not fed back into the ED database, thus severing connections and leaving ED database patients in isolation with redundant, temporary PAS IDs.

	• Aside from the one-way transfer of information from the PAS database to the ED database, no database has the functionality to store other System IDs.

Factor	Description
Database access	• Staff are restricted to accessing the CIS within their departments. **UCC staff** • Both administrative and healthcare staff can access the UCC Database. • Healthcare staff can also view results on the Radiology and Pathology databases. **ED staff** • Both administrative and healthcare staff can access the ED Database. • Healthcare staff can also view results on the Radiology and Pathology databases. • Administrative staff can access the UCC Database to register patients.
Paper use	**UCC Diagnostic test requests** • The UCC database does not have the functionality to request diagnostic tests for patients electronically. • Thus, UCC healthcare staff use paper, by printing a diagnostic request form, for the patient to take to the ED receptionist who creates an EPR within the PAS database and handwrites the PAS ID onto the diagnostic request form, to enable the diagnostic test to be undertaken. • The patient returns to the UCC, and the healthcare professional uses the PAS ID to access results within the Pathology or Radiology databases. • The paper with the PAS ID is destroyed, and the PAS ID is not recorded within the UCC database as there is no designated field. **UCC patient requires referral to ED (Red stream)** • In some cases, patients may require referral to ED services, following an initial assessment by a UCC GP Streamer. • As the UCC database is 'isolated' from other systems within the emergency setting, the GP streamer will print a paper referral form, containing a short summary and hand this to the patient so that they may register at ED reception. • The ED receptionist then follows the registration process as per Figure 5.1 (page 203) **ED casualty cards** • Within ED, although an EPR is created within the ED Database, clinical staff print a paper template 'Casualty Card' that contains Episode ID, Patient ID and other details transcribed by the ED receptionist. • All details about a patient's episode of care are handwritten onto the paper Casualty Cards and only mandatory details transcribed within the ED Database. • Upon conclusion of the patient episode, the Casualty Cards are scanned, an electronic file is created and attached to the EPR within the ED Database.

Factor	Description
Transition time	**Patient queuing time between service referrals** • All UCC walk-in patients initially queue to be seen by UCC reception. • The GP streaming process takes time, and those UCC patients requiring referral to ED (Red stream) must queue again to register with the ED receptionist, requiring additional time. • Although the 'transition' from UCC to ED takes additional patient time, it is not recorded within any system, yet it contributes to the total time a patient stays within the department and is affected by busy periods.
Admitted patients	• Patients requiring emergency admission are sent to IPC • Such patients have already been allocated a unique PAS ID and their details and diagnostic test results already recorded within the PAS system. • Hospital wards and specialties can use pre-existing information to further aid in the management of patients.
Duplication of patients	• Red stream patient details are being recorded within both the UCC and ED databases, respectively. However, for analysis purposes, if both UCC and ED services count the same patients, this will lead to an overestimation in the total number of attendances throughout urgent and emergency care. • A prior agreement between the UCC and ED services is that the UCC will exclude all red stream patients from its total attendance and analysis. • However, there are some cases where patients that arrive via ambulance, are triaged in ED, but their condition is found to be more appropriate for the UCC; hence, are referred to the UCC for treatment. In spite of this, there is no detailed process to reconcile nor verify the actual correct numbers.

5.2.2 Use case diagram

Use case name	Patient injury	
Description	This use case describes the process of a new patient presenting to the UCC seeking treatment for a foot injury.	
Precondition	Patient must have a minor injury requiring X-Ray.	
Postcondition	Patient injury is investigated, diagnosed and treatment provided.	
Primary actors	Patient	
Other actors	**Boundaries**	Outside hospital, Inside hospital.
	Systems	UCC Database, ED:PAS Database, Radiology Database
	Human actors	UCC Receptionist, GP Streamer, Emergency Nurse Practitioner, ED Receptionist, Radiographer/Radiologist.
	Non-human actors	UCC Database provider, NHS Personal Demographic Service, Hospital PAS Database provider, Radiology Database Provider

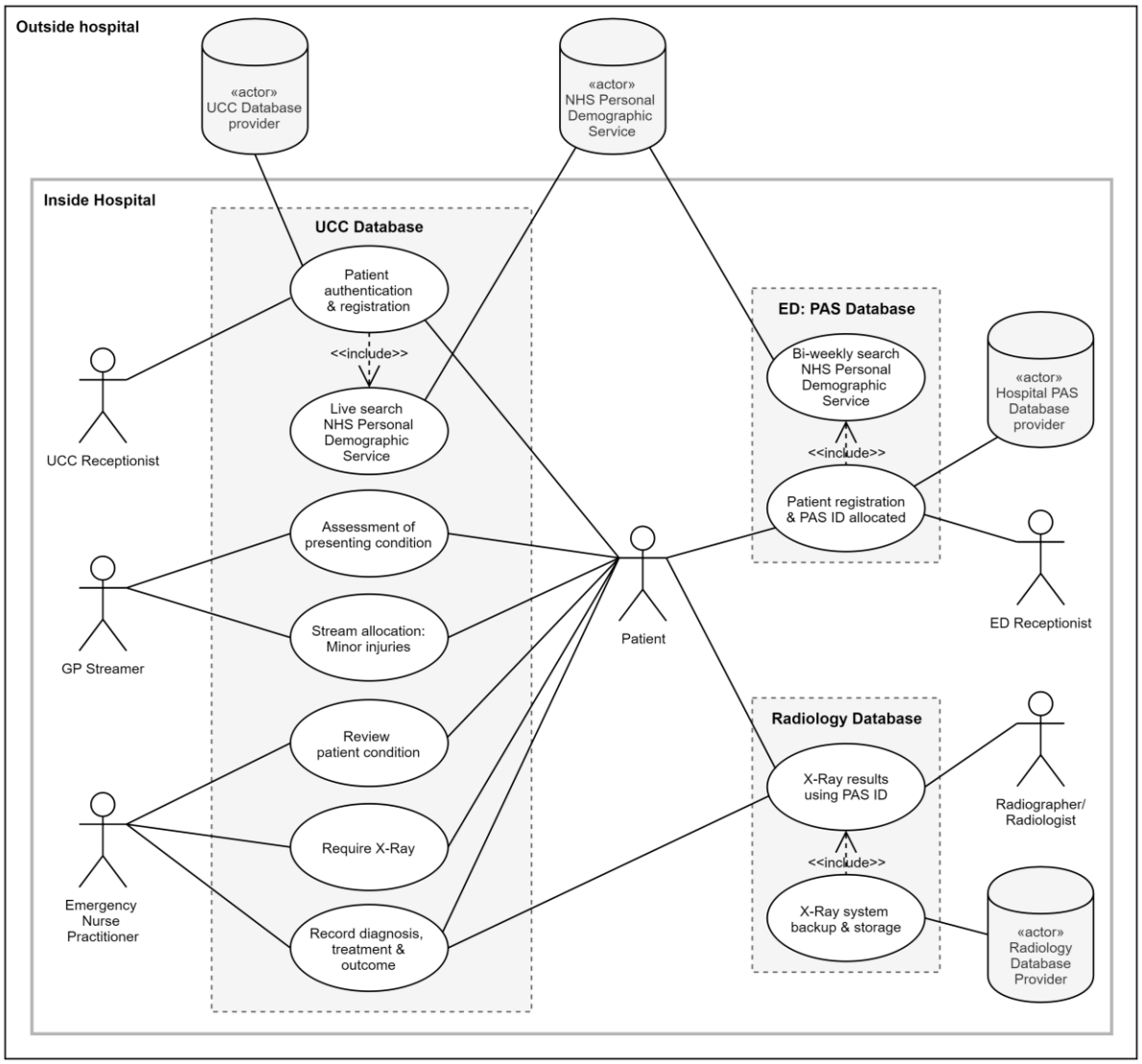


Figure 5.2 Use case diagram highlighting databases encountered during patient episode

Table 5.2 Interactions between actors, based on Use case diagram (Figure 5.2)

Use cases	Description of interactions between actors
> **Within UCC Database** • Patient authentication and registration <<includes>> • Live search NHS Personal Demographic Service	Patient arrives at the Hospital and UCC receptionist registers details within UCC Database, a new EPR is created which includes patient's details being verified on the NHS Personal Demographic Service (NHS PDS) via a live secure internet connection (outside the hospital). Verified details are transmitted to the UCC Database, including NHS number, and registration completed. The UCC Database provider maintains UCC Database storage and backup.
> **Within UCC Database** • Assessment of presenting condition • Stream allocation: Minor Injuries	The GP streamer performs an initial assessment of the patient's injury and allocates the patient to the Minor injuries stream. Patient then waits in reception.
> **Within UCC Database** • Review patient condition • Require X-Ray	Patient seen by the Emergency Nurse Practitioner (ENP) for a review of their injury/condition. ENP decides that an X-Ray is required and prints out request form and hands to patient.
> **Within ED PAS Database** • Patient registration PAS ID allocated <<includes>> • Bi-weekly search NHS PDS	Patient then queues at ED reception, and ED receptionist registers details within PAS Database, creates new EPR. This includes patient's details being verified against a Local NHS PDS repository (downloaded bi-weekly via a secure internet connection, from outside the hospital). The PAS ID is handwritten onto printed X-Ray request form and given to the patient. The EPR is concluded within the PAS Database. The Hospital PAS Database provider maintains PAS Database storage and backup.
> **Within Radiology Database** • X-Ray results using PAS ID <<includes>> • X-Ray system backup & storage	Patient presents to the Radiology department, hands Radiographer the X-Ray request form. The Radiographer uses the PAS ID to identify patient and takes X-Rays of the patient's foot, which are digitally stored within the Radiology Database for the Radiologist to assess. The request form is returned to the patient to take to the UCC ENP. The Radiology Database provider maintains Radiology Database storage and backup.
> **Within UCC Database** • Require X-Ray • Record diagnosis, treatment & outcome	Patient returns to UCC, informs UCC receptionist that X-Ray has been completed and then waits in a waiting room. The ENP calls the patient (post X-Ray) who provides the request form along with the PAS ID. The ENP access the Radiology Database using the PAS ID to review the X-Ray and views accompanying comments. As the patient was found to have a fracture, the patient receives appropriate treatment and is then discharged from the UCC. The ENP concludes the EPR within the UCC Database. The UCC Database provider maintains UCC Database storage and backup.

5.2.3 Heterogeneity amongst CIS

Table 5.3 (page 209) provides a summary of the findings from the Activity diagram (page 203) and the Use Case diagram (page 207) highlight identified interrelationships and levels of interoperability or differences between each CIS.

Table 5.3 Heterogeneity amongst urgent & emergency care CIS

	Department/service		
CIS detail	**UCC** (GP-led Urgent Care Centre)	**ED** (Emergency Department)	**IPC** (Inpatient Care)
Database	UCC DB	ED DB	IPC DB
Stores	Attendance episodes	Attendance episodes	Admission spells
Record Structure	Patient record	Patient record	Health record
Medical record structure	Chronological episodes of attendance	Chronological episodes of attendance	Patient-oriented Admission spells
Identification numbers	UCC Episode ID UCC Patient ID NHS number	ED Episode ID ED Patient ID Temporary PAS ID NHS number (via PAS)	PAS Patient ID PAS admission spell ID NHS number
Primary identifier	Episode ID	Episode ID	Patient ID
NHS number storage	Direct link to NHS Personal Demographics Service	No link to NHS PDS, Updated via electronic message from ICHIS	Updates twice a week from NHS Personal Demographics Service
Storage	Web server external	Local hospital server	Local hospital server
Interoperability with other CIS	None, completely stand alone	Receives one-way message from PAS DB	One-way, sends messages to ED DB

The UCC database was stored on an external web server, whilst the ED and IPC databases were stored on hospital premises. This could explain why the UCC database was completely isolated from the ED and IPC databases. Additionally, there was only one-way transmission of data from the IPC to ED database. Within the ED database, the allocation/storage of a temporary PAS ID which ultimately becomes redundant and is not updated, will only serve to isolate such records.

The EMR structure varied between the three CIS, both the ED and UCC databases were based on chronological records emphasising date/time of attendance, whereas the IPC database was patient-oriented. The NHS number does not appear to be updated regularly within the ED database, so its

capture may be sparse. These findings help to gauge the level of heterogeneity and the hurdles that must be overcome to enable data linkage and integration.

The data dictionaries for each of the three departmental CIS provided by the vendors revealed different underlying relational models, table structures and naming conventions. One dictionary stored outdated tables that did not reflect the modifications or customisations in the live CIS, whilst other dictionaries contained more intuitive schematic diagrams.

The use of aliases and pseudonyms for naming conventions, resulted in discrepancies between the dictionary and implemented CIS. For example, the degree of heterogeneity between the three CISs, with respect to the naming convention for the Patient Identifier within each CIS data dictionary:

- **UCC-Database:** PatientRef
- **ED-Database:** GH2_Patient_Godsnumber_View
- **IPC-Database:** PKEY

Additionally, the process of domain understanding discovered that heterogeneity amongst the UCC, ED and IPC databases was inherent, and respective data dictionaries confirmed that the patient identification number used by each CIS were completely heterogeneous and would not support data linkage, i.e. each departmental CIS automatically generated patient IDs in different formats:

- **UCC-Database** (alphanumeric): 123a4567-8b9c-01d2-3456-e7890123f45g
- **ED-Database** (numeric): 1234567
- **IPC-Database** (alphanumeric): 123456A

Data linkage is straightforward when a unique identification number such as an NHS number is readily available. [202] However, for this study, the use of the NHS number for data linkage was not an option as the ED database did not link to the NHS Personal Demographics Service, hence, the NHS number was not being recorded within all patient records. Therefore, one may anticipate poor DQ (completeness) of the NHS number, which poses a major limitation for prospective record linkage endeavours using the NHS number to link datasets.

Consequently, another means to identify a patient uniquely may be required, and the literature review identified a data linkage strategy using only quasi-identifiers, i.e. patient demographic information, forename, surname, sex, date of birth and postcode. [202, 217, 220, 221] Although this

is a more challenging approach, such patient information will be routinely recorded and readily available within each of the three CIS. Furthermore, the author was granted read-only permission to use the front interface to create custom queries/reports within each of the three CISs, for data extraction purposes.

5.3 Semantic interoperability

Addresses objective 3: To apply techniques that enable interoperability between the clinical information systems in the emergency care setting.

5.3.1 Ontologies

A person is an individual human being about whom information is maintained; i.e. a PERSON may be a PATIENT attending a hospital. [311] Therefore, to eliminate data heterogeneity, the concept of a 'Patient' with attributes: Name, Gender, Birth date, Address and Identifier was created. [149, 311]

Table 5.4 Person attributes, adapted from NHS Digital (2019).

Attribute	Description
Patient	A PATIENT is a PERSON with a specific disease or condition who receives treatment from a Health Care Provider
Identifier	An identifier, other than a name, which identifies a PERSON
Birth date	The date on which a PERSON was born or is officially deemed to have been born
Person name	• The unique identifier for a specific and ordered combination of words and titles by which a PERSON may be known. • A full name comprised of one or more separate PERSON NAMES in sequence. • The PERSON NAME must contain at least one PERSON NAME of the type 'Person Family Name'.
Address	The identification of a place of relevance to a PERSON.
Postcode	The code assigned by Royal Mail to identify postal delivery areas across the United Kingdom.

Source: NHS Digital, 2019 [311]

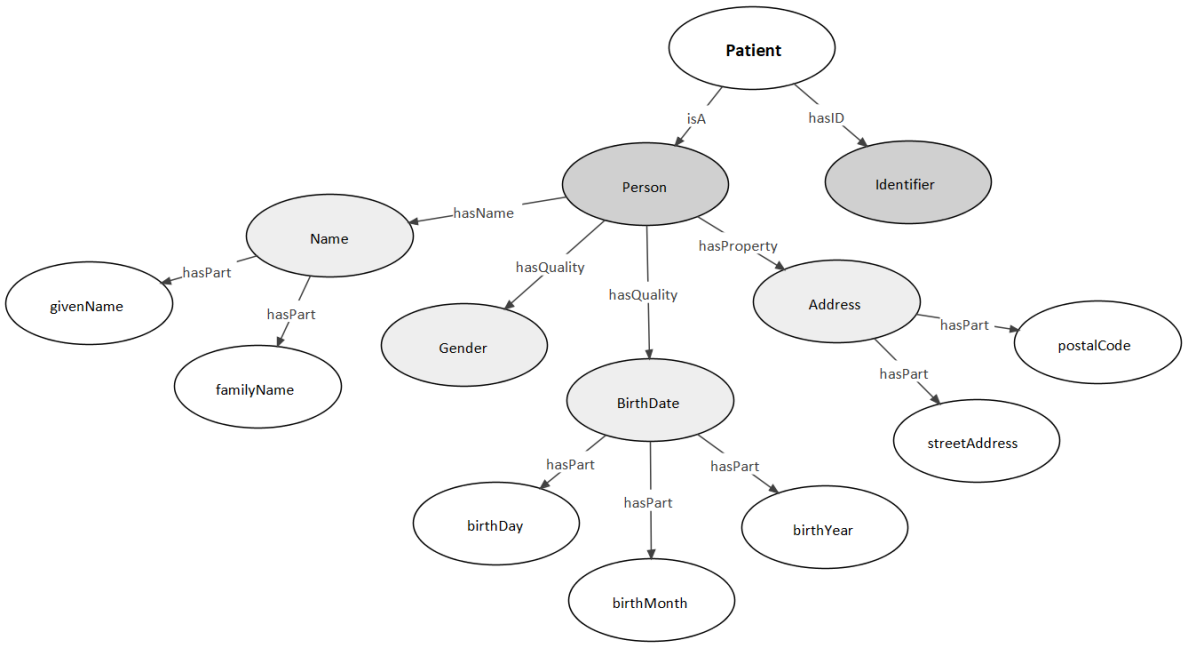


Figure 5.3 Patient ontology model

Figure 5.3 (page 212) illustrates the patient ontology model, with descriptions of attributes as per Table 5.4. The concept of a patient as a person with specific attributes exists, independent of any CIS. Thus, a single ontology approach, [157] was used to create a patient ontology with specific properties common to all three CISs, and permitted the identification of common data elements across the three departmental CISs, regardless of the heterogeneity identified previously amongst naming conventions.

5.3.2 Standardisation of data elements

The ontology approach utilised the 'concept' of a patient with specific attributes that were cross-referenced with the UCC, ED and IPC data dictionaries, respectively. Having referred to HL7 FHIR [160] and the NHS Data Dictionary [311] for guidance on standardisation, specific data elements required for data linkage and integration were identified and standardised via application of a common naming convention, format and coding classification.

Furthermore, using the reporting tools within each respective CIS permitted the creation of bespoke reports, which specified the data elements that were to be extracted, and permitted the allocation of an 'Alias' to ensure data elements from all three CIS would be semantically interoperable.

Table 5.5 (page 213) provides an overview of the heterogeneity amongst naming conventions as recorded within the respective data dictionary dictionaries from each CIS. Additionally, the 'Alias' column serves to provide one standardised naming convention for each data element.

Table 5.5 Standardisation of naming conventions

		Original database naming conventions		
Data elements	**Alias**	**UCC**	**ED**	**IPC**
Identification numbers				
Episode ID	**EpId**	CaseRef	GH2_EpisodeID	DistrictSpellNumber
Patient ID	**PId**	PatientRef	GH2_Patient_ Godsnumber_View_ psi_system_id	PKEY
NHS Number	**Nhs**	NHSNumber	GH2_NHS_numbers_ View_psi_system_id	NHSNumber
Quasi Patient identifiers				
Surname	**Sn**	Surname	GH2_Patient_pat_surname	Surname
Forename	**Fn**	Forename	GH2_Patient_pat_forename	Forename
Sex	**Sx**	Gender	GH2_Lookups_Sex_ Lkp_Name	Gender
Date of birth	**Dob**	DateOfBirth	GH2_Patient_DOB	DateOfBirth
Postcode	**Pc**	Postcode	GH2_pat_details_view_ add_postcode	PostCodeOn Admission
Episode dates/times				
Arrival Date	**ArrD**	ActiveDateTime	GH2_Arrival_Date	AdmissionDate
Arrival Time	**ArrT**	ActiveDateTime	GH2_Arrival_Time	AdmissionTime
Depart Date	**DepD**	LastConsultationEnd	GH2_Leave_Date	DischargeDate
Depart Time	**DepT**	LastConsultationEnd	GH2_Leave_Time	DischargeTime
Episode details				
Hospital site	**HospSite**	Site	GH2_Site	SiteName
Clinical stream	**Strm**	FinalStream	GH2_First_Stream_Stream	-
Outcome	**Outc**	Informational Outcomes	GH2_Lookups_dischout_ Lkp_Name	ichisdescrip2

5.4 Syntactic interoperability

Addresses objective 4: To select data warehouse architecture appropriate for housing healthcare data extracts.

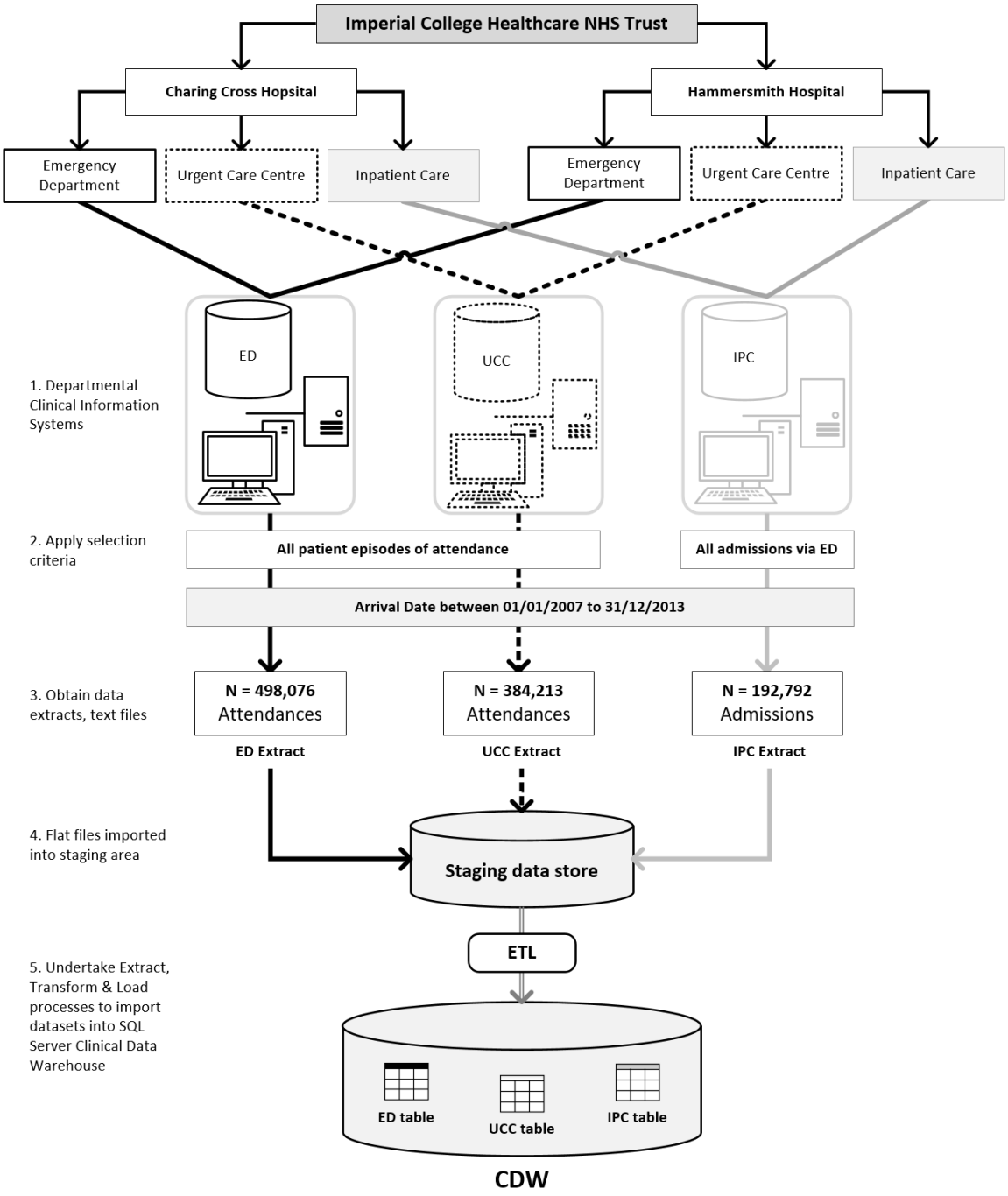


Figure 5.4 Data extraction and import into the clinical data warehouse

Each of the three CISs were heterogeneous, operating in isolation from one another. As there was no syntactic interoperability between the three departmental CIS, the Data-driven approach, as per Inmon's DW Architecture [173] was utilised.

Within the context of this study, the Inmon approach consisted of creating bespoke reports with accompanying queries that enabled the identification and selection of specific data elements and application of time frame criteria within each respective departmental CIS. Moreover, all three CISs were configured to record information across both CXH and HH sites, this feature enabled the simplification of extraction queries.

Figure 5.4 (page 214) provides an overview of the process in which three data extracts were obtained, and following import into a secure staging area, the extracts were processed and transformed to standardise naming conventions and data element formats/properties to facilitate semantic interoperability.

Following the import of the three data extracts into the CDW, an assessment of data quality would be required, and an appropriate record linkage strategy selected/applied prior to the creation of a Linked dataset which could then be used for further analysis; thus fulfilling the remaining objectives of this study.

5.5 Data quality

5.5.1 Data quality assessment

Addresses objective 5: To determine if data are fit for linkage purposes by assessing the quality of data.

An assessment of Data Quality (DQA) for each data element is presented in the following tables.

5.5.1.1 DQA: Episode ID

Table 5.6 Episode ID data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Episode ID	Attendance episode	Attendance episode	Admission spell
Original name	CaseRef	GH2_EpisodeID	DistrictSpellNumber
Example format	12345abc-a1bc-1234-a123-123a456789b0	AB-12-345678-9	Incremental 1 + 1234567
Type	String	String	Number
Total records (N)	384215	498081	192792
Missing (n)	0	0	0
Completeness	384215 (100.0)	498081 (100.0)	192792 (100.0)
Uniqueness	384215 (100.0)	498081 (100.0)	192792 (100.0)
Character length (n)	36	14	6 or 7
Inconsistent (n)	0	1	1
Consistency	384215 (100.0)	498080 (99.9)	192791 (99.9)
Validity	384215 (100.0)	498081 (100.0)	192792 (100.0)

Referring to Table 5.6 (page 216) within the UCC and ED datasets, the episode ID uniquely identifies one episode of attendance, and within the IPC dataset, it uniquely identifies one admission spell. Hence, the episode ID could be thought of as the primary key for each dataset, as there were no duplicate values, and the overall completion was 100% within each dataset.

However, it was discovered that each episode ID was, in fact, heterogeneous and specific per parent CIS, for example, UCC dataset (36 character string), ED dataset (14 character string) and IPC dataset (6 or 7 character number) are inherently different. Consequently, the unique structure would make it impossible to link any of the datasets using DRL techniques.

For the ED and IPC datasets, only one value was found as being inconsistent with other data elements, due to different character length, but these remained valid entries. The total number of records will serve as the denominator for data quality checks on the remaining data elements.

Moreover, as expected the Episode ID did not contain any duplicate values, with each CIS generating a new Episode ID for each new patient attendance or spell of admission.

5.5.1.2 DQ: Patient ID

Table 5.7 Patient ID data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	PatientRef	GH2_Patient_ Godsnumber_View_ psi_system_id	PKEY
Example format	123a4567-8b9c-01d2-3456-e7890123f45g	123456A	1234567
Type	String	String	Number
Count (n)	384215	498074	192792
Missing (n)	0	7	0
Completeness	384215 (100.000)	498074 (99.999)	192792 (100.000)
Uniqueness	229886 (59.833)	264415 (53.087)	108480 (56.268)
Duplication level	1.671	1.884	1.777
Character length (n)	36	3 to 9	7
Inconsistent (n)	0	20	0
Consistency	384215 (100.000)	498054 (99.995)	192792 (100.000)
Validity	384215 (100.000)	498074 (99.999)	192792 (100.000)

Within Table 5.7 (page 217), the patient ID uniquely identified individual patients within each dataset, and as per the data dictionary, each ID was inherently different. When looking at the ED and IPC patient IDs, although both originate from the IPC database, the temporary patient ID transferred from IPC to ED database (when a patient first registers) ends with the suffix ‘A’, there were found to be 20 records that were inconsistent with this format. Additionally, the ED dataset contained seven records with no recorded patient ID.

Moreover, upon conclusion of the patient episode, the IPC database team retrospectively review/allocate a permanent numeric patient ID within the IPC database; thus severing any links with the ED system. Ultimately, each patient ID has a unique structure, and therefore does not support linkage using DRL techniques. Overall the completion rate was high, with a minimum of 99.99% within ED dataset.

5.5.1.3 DQA: NHS Number

Table 5.8 NHS number data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	NHSNumber	GH2_NHS_numbers_ View_psi_system_id	NHSNumber
Example format	123_456_7890	123_456_7890	1234567890
Type	String	String	Integer
Count (n)	301730	228294	182726
Missing (n)	82485	269787	10066
Completeness	301730 (78.532)	228294 (45.835)	182726 (94.779)
Uniqueness	164895 (54.650)	86346 (37.822)	99021 (54.191)
Duplication level	2.330	5.768	1.947
Character length (n)	12	12	10
Inconsistent (n)	0	0	0
Consistency	301730 (78.532)	228294 (45.835)	182726 (94.779)
Validity	301730 (78.532)	228294 (45.835)	182726 (94.779)

Referring to Table 5.8 (page 218), the method by which the NHS number is allocated within each CIS varies as does the completion rate between each dataset. As expected, the ED dataset had the lowest recording for NHS number (approximately 45.8%), which supports the finding that the ED database is completely stand-alone/isolated and does not link to the NHS Patient Demographics Service (NHS-PDS). Interestingly the NHS number within the ED database is only populated when a patient presents for the second time, and it receives an updated one-way message from the IPC database, containing the NHS number.

Conversely, the NHS number completion for both the UCC dataset (78.5%) and IPC dataset (94.7%) was much higher, as both have connections to the NHS-PDS, with the UCC database updating instantaneously for new patient registrations. Additionally, IPC scored the highest completeness as it was the hospital's central PAS system, and had a dedicated team updating and maintaining it retrospectively.

Although the NHS number was standardised to an integer format across all three datasets and would readily link using DRL techniques, the completeness of the ED dataset would be the rate-limiting step. Moreover, all of the NHS numbers within each dataset were checked using the NHS validation algorithm (detailed on page 108), and all were found to be valid.

5.5.1.4 DQA: Patient surname

Table 5.9 Patient surname data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	Surname	GH2_Patient_pat_surname	Surname
Type	String	String	String
Count (n)	384215	498081	192790
Missing (n)	0	0	2
Completeness	384215 (100.000)	498081 (100.000)	192790 (99.999)
Uniqueness	88259 (22.971)	90825 (18.235)	42287 (21.934)
Duplication level	4.353	5.484	4.559
Character length (n)	1 to 40	1 to 33	2 to 16
Inconsistent (n)	26297	56033	31615
Consistency	357918 (93.156)	442048 (88.750)	161175 (83.600)
Invalid (n)	21	590	73
Validity	384194 (99.995)	497491 (99.882)	192717 (99.961)

5.5.1.5 DQA: Patient forename

Table 5.10 Patient forename data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	Forename	GH2_Patient_pat_forename	Forename
Type	String	String	String
Count (n)	384209	498056	192787
Missing (n)	6	25	5
Completeness	384209 (99.998)	498056 (99.995)	192787 (99.997)
Uniqueness	40976 (10.665)	41935 (8.420)	20801 (10.790)
Duplication level	9.377	11.877	9.268
Character length (n)	1 to 25	1 to 14	1 to 12
Inconsistent (n)	2542	1741	448
Consistency	381667 (99.337)	496315 (99.645)	192339 (99.765)
Invalid (n)	15	250	23
Validity	384194 (99.995)	497806 (99.945)	192764 (99.985)

Within all three datasets, both the surname (Table 5.9, page 219) and forename (Table 5.10, page 219), data elements were above 99% complete. While cleaning the respective reference tables, 'Test' patient entries were identified, along with other spurious entries including 'Anony', 'Test', 'Dummy', 'Dustbin', 'aehh'. Identification of such entries was not a straightforward task as evidently

dummy data had been entered when testing the CISs before they went live, with transcribers using random names, including spaces and special characters. Ideally, these should have been removed from the live system. In total, there were two episodes within the UCC dataset and five within ED dataset that required exclusion.

All text where 'Unknown' had been recorded, were set to Null to prevent any false associations. In seven cases within UCC dataset, the name 0'Surname (Zero'Surname) was identified, and these records were edited manually, e.g. 0Surname, 2Forename, 1234-5678-9012Surname, DOB placed instead of a name. Additionally, neonates were given the forename 'baby', and the surname was recorded as the family name.

The forename element contained much more variation, approximately 10% overall uniqueness when compared to the surname element, approximately 21% uniqueness. Hence, the chances of two persons have similar surnames was higher than the chances of having similar forenames. In addition, names were very much dependent on a person's ethnicity or country of origin.

Consistency was comparatively higher for the forename (approximately 99% overall) compared to the surname (approximately 88%) this may be due to the presence of characters, e.g. hyphens, symbols or extra spaces within the surname data elements. Of the three datasets, ED contained the highest number of invalid forenames (n=250) and over double the number of invalid surnames (n=590), but due to the greater number of episodes (denominator), the overall validity was 99.9%.

Moreover, as the name data element is purely free text, this is where the most transcribing errors can occur. Such errors may be indirectly affected by how busy a department is; the keyboard skills of the person transcribing and even the accent of the patient in pronouncing their name and how the spelling is interpreted, e.g. Green or Greene.

5.5.1.6 DQA: Patient sex

Table 5.11 Patient sex data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	Gender	GH2_Lookups_Sex_Lkp_Name	Gender
Example format	F, M	Female, Male	F, M
Type	String	String	String
Count (n)	384215	497902	192792
Missing (n)	0	179	0
Completeness	384215 (100.000)	497902 (99.964)	192792 (100.000)
Uniqueness	4 (0.001)	3 (0.001)	3 (0.002)
Character length (n)	1	4 to 7	1
Inconsistent (n)	80	6	3
Consistency	384135 (99.979)	497896 (99.963)	192789 (99.998)
Invalid (n)	82	8	3
Validity	384133 (99.979)	497894 (99.962)	192789 (99.998)

Within Table 5.11 (page 221), the coding for data element 'Sex' was found to vary amongst the three CIS, with both UCC and IPC opting to use the initials, and ED the full text. Inconsistent characters such as 'Unknown', 'N', 'U' or '?' were identified and set to Null. The completion and validity of the sex data element was greater than 99.9% across all three datasets. Although, this data element tends to be a 'drop-down list (to restrict choices); some CISs permit overwriting of what is listed.

5.5.1.7 DQA: Patient date of birth

Table 5.12 Patient date of birth data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	DateOfBirth	GH2_Patient_DOB	DateOfBirth
Example format	YYYY-MM-DD	YYYY-MM-DD	YYYY-MM-DD
Type	Date	Date	Date
Minimum value	1888-08-04	1800-01-01	1880-01-01
Count (n)	384201	497825	192765
Missing (n)	14	256	27
Completeness	384201 (99.996)	497825 (99.949)	192765 (99.986)
Uniqueness	31128 (8.102)	33207 (6.670)	28903 (14.994)
Character length (n)	19	19	19
Inconsistent (n)	0	0	0
Consistency	384201 (99.996)	497825 (99.949)	192765 (99.986)
Invalid (n)	32	1112	200
Validity	384169 (99.988)	496713 (99.725)	192565 (99.882)

Referring to Table 5.12 (page 221), in cases where the date of birth data element was unknown, some hospital CISs record a default date of either '01/01/1900' or '01/01/1901'. [342] Weber et al. (2012), [343] have corroborated this finding, reporting that these dates were being used as placeholder values for unknown dates of birth, and opted to exclude such dates.

Across the UCC, ED and IPC datasets, different types of 'unknown' dates were identified, including: 1800-01-01, 1860-01-01, 1880-01-01, 1888-01-12, 1888-08-04, 1900-01-19, 1900-02-20, 1900-12-19, 1901-01-01, 1901-01-19, 1901-07-08, 1901-10-05, 1901-10-30. For the purposes of this research, birth dates less than or equal to '1901/01/01' were set to null. Additionally, splitting the birth date into three separate components: Birth Year, Month and Day facilitated further checks for ranges between 1 month and 12 months; and 1 day and 31 days.

Within the ED dataset, one case was identified where the date of birth was greater than the arrival date; this specific case was cross-referenced, from the ED dataset to other datasets, using the patient's NHS number and the correct date of birth identified and ED dataset updated, whilst informing the ED database administrator. Some additional cases were identified whereby the date of birth and arrival date were equal, and further investigation revealed this to be indicative of neonates, i.e. birthday equals the arrival date. Moreover, the datasets revealed that at least 99.7% of birth dates were valid.

5.5.1.8 DQA: Postcode

Table 5.13 Patient postcode data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	Postcode	GH2_pat_details_view_add_postcode	PostCodeOnAdmission
Type	String	String	String
Count (n)	380292	481049	191626
Missing (n)	3923	17032	1166
Completeness	380292 (98.979)	481049 (96.580)	191626 (99.395)
Uniqueness	53543 (14.079)	69713 (14.492)	35411 (18.479)
Duplication level	7.176	7.145	5.444
Character length (n)	2 to 14	1 to 9	8
Inconsistent (n)	471	2877	2053
Consistency	379821 (98.856)	478172 (96.003)	189573 (98.330)
Invalid (n)	62	12	0
Validity	380230 (98.963)	481037 (96.578)	191626 (99.395)

The postcode data element (Table 5.13, page 222) across all three datasets identified between 95.9% (in ED) to 98.8% (in UCC) as UK postcodes. Additional mapping to the NHS postcode directory, as described in the methods section, helped to provide further clarity and aided validation and cleaning processes. For example, some postcodes with a prefix of ZZ were correctly identified as being overseas patients, i.e. between 0.6% (in ED) to 1.1% (in IPC). However, in other cases, the text 'unknown' or 'ZZ99' or another derivative, had been recorded; which remained unmapped; this accounted for less than 0.02% from across all three datasets.

5.5.1.9 DQA: Arrival Date/Time

Referring to Table 5.14 (page 223) and Table 5.15 (page 224) within the ED and IPC datasets, the date and time elements were merged to create a new date/time element, which would facilitate date difference calculations. The arrival date/time in UCC and IPC showed 100% completion, whereas ED had 55 missing values resulting in 99.9% being recorded. Additionally, for the IPC dataset, the completeness for depart date was 99.9%, and not recorded for some records (n=35), which may be due to active patient episodes, at the time of obtaining the data extracts.

Table 5.14 Episode arrival date/time data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	ActiveDateTime	GH2_Arrival_Date GH2_Arrival_Time	AdmissionDate AdmissionTime
Example format	YYYY-MM-DD HH:MM	[YYYY-MM-DD] [HH:MM]	[YYYY-MM-DD] [HH:MM]
Type	Merged [Date + Time]	Separate [Date][Time]	Separate [Date][Time]
Count (n)	384215	498026	192792
Missing (n)	0	55	0
Completeness	384215 (100.000)	498026 (99.989)	192792 (100.000)
Uniqueness	383637 (99.850)	464253 (93.219)	183460 (95.160)
Duplication level	1.002	1.073	1.051
Character length (n)	19	19	19
Inconsistent (n)	0	0	0
Consistency	384215 (100.000)	498026 (99.989)	192792 (100.000)
Invalid (n)	2	2	0
Validity	384213 (99.999)	498024 (99.989)	192792 (100.000)

5.5.1.10 DQA: Depart Date/Time

Table 5.15 Episode depart date/time data quality summary

	Dataset (n, %)		
Detail	**UCC**	**ED**	**IPC**
Total episodes (N)	**384215**	**498081**	**192792**
Naming convention	LastConsultationEnd	GH2_Leave _Date GH2_Leave_Time	DischargeDate DischargeTime
Example format	YYYY-MM-DD HH:MM	[YYYY-MM-DD] [HH:MM]	[YYYY-MM-DD] [HH:MM]
Type	Merged [Date + Time]	Separate [Date][Time]	Separate [Date][Time]
Count (n)	384215	498026	192757
Missing (n)	0	55	35
Completeness	384215 (100.000)	498026 (99.989)	192757 (99.982)
Uniqueness	383637 (99.850)	464253 (93.219)	183460 (95.177)
Duplication level	1.002	1.073	1.051
Character length (n)	19	19	19
Inconsistent (n)	0	0	0
Consistency	384215 (100.000)	498026 (99.989)	192757 (99.982)
Invalid (n)	2	2	0
Validity	384213 (99.999)	498024 (99.989)	192757 (99.982)

5.5.1.11 DQA: Hospital Site

The naming convention used for recording hospital site varied between the three datasets:

- **UCC-Database:** CXH, HH
- **ED-Database:** CX, HH
- **IPC-Database:** CHARING CROSS HOSPITAL, HAMMERSMITH HOSPITAL

As the analysis to be undertaken for Objectives 7 and 8 in will be per site, this data element was standardised to CXH and HH for all three datasets. Additionally, there was found to be 100% completeness within each dataset.

5.5.1.12 DQA Summary

Figure 5.5 (page 225) provides an overview/comparison of all three datasets, illustrating the level of completeness, post validation/cleaning, for the specific data elements, as assessed in this section. With the exception of the NHS number, the overall completeness of direct patient identifiers are above 96%, which strongly implies that these identifiers will be appropriate candidates for subsequent record linkage purposes.

For example, the forename will not usually change throughout a person's lifetime; a surname may change as a result of marriage; the date of birth and gender will not change; postcodes are subject to change, but people tend not to move address too frequently. Hence, such patient quasi-identifiers may also facilitate the unique identification of a patient across the three datasets.

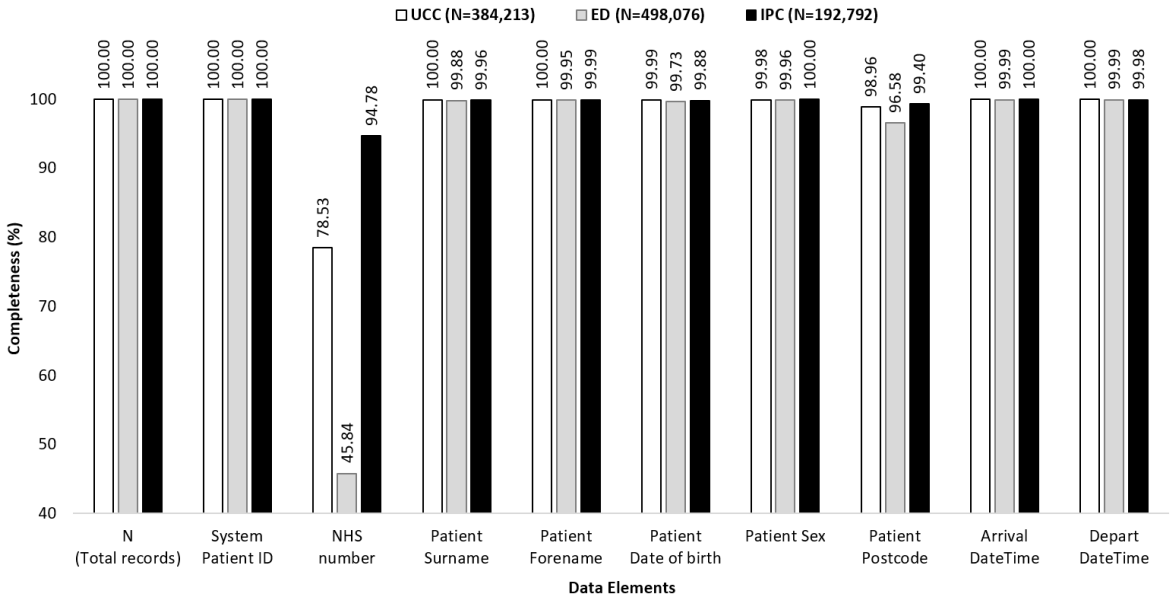


Figure 5.5 Data quality dimension: Completeness for selected data elements

5.6 Data linkage

Objective 6: To differentiate the effect of deterministic and non-deterministic data linkage strategies with respect to the numbers of correctly matched records.

5.6.1 Hybrid (DRL & PRL) linkage

5.6.1.1 Unique patient identifier

The unique identification of a patient was compounded by the number of ways in which a patient could be identified within and across each distinct dataset due to the presence of multiple heterogeneous patient identifiers and their underlying complex interrelationships.

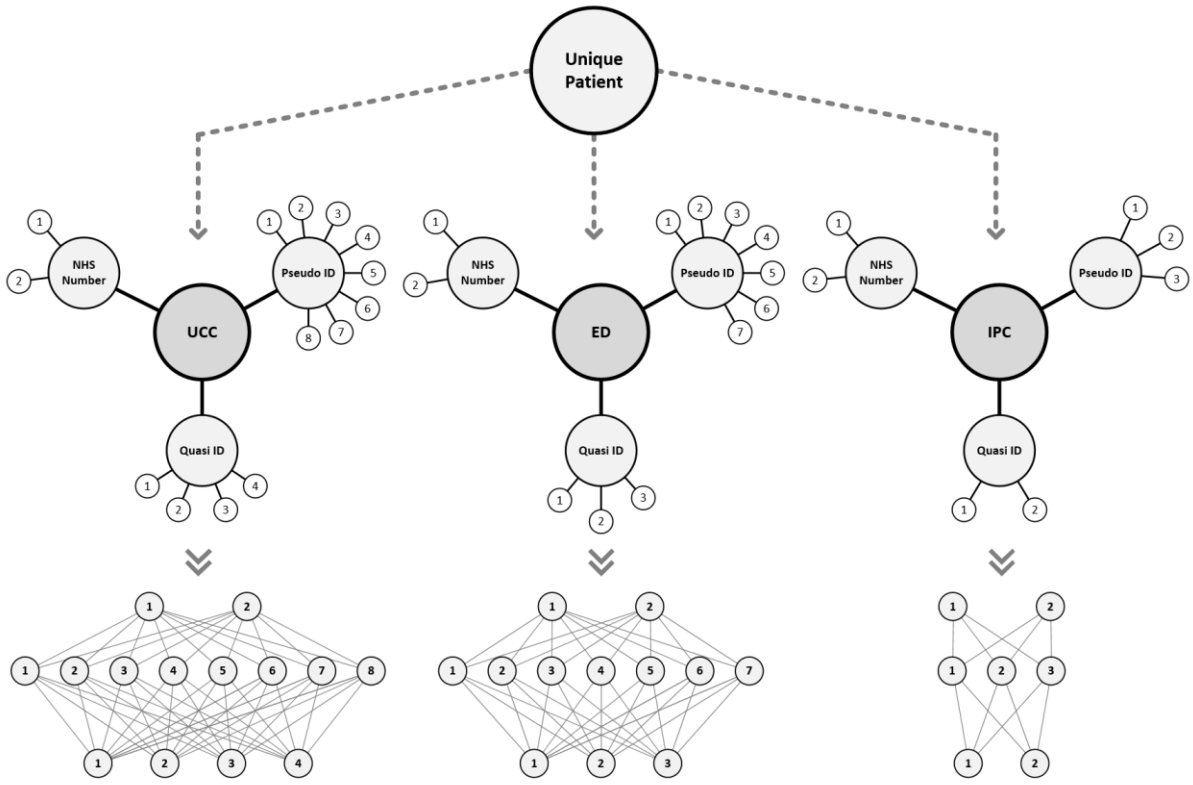


Figure 5.6 Hidden complex interrelationship between relative CIS IDs for a unique patient within each dataset.

For example, referring to Figure 5.6 (page 226) and taking the UCC dataset perspective, a unique patient was found to have between zero, one or two NHS numbers, one to eight different Pseudo ID's (System Patient ID) and between one and four different Quasi-ID's (variations in Surname,

Forename, Date of birth, Sex). The ED dataset showed marginally less variability than the UCC dataset, whilst the IPC dataset was found to have the least variability in Pseudo ID.

Although the NHS number was common across all three datasets, its level of completeness was very low, approximately 46 per cent. In addition, the Pseudo ID, which represented respective CIS patient IDs, was highly heterogeneous and did not support patient identification across multiple datasets. However, as the Quasi-ID was based on patient demographic details, it was able to readily identify the same patient across multiple datasets.

5.6.1.2 Data element variants per ID

For each unique ID, the number of variants per data element was cross-referenced to determine the amount of variability, e.g. one Pseudo ID was associated with nine postcode variants, within the IPC dataset.

Referring to Figure 5.7 (page 228), to uniquely identify a patient across the three datasets, there were three ID's available, i.e. the Pseudo ID, NHS number and the Quasi-ID. Viewing the data from the Pseudo ID or the NHS number perspective revealed variation in the number of surnames and forenames, i.e. a unique patient might have between one to four variants on names.

Similarly, some patients were recorded as both male and female; whilst others were found to have between one to three variants on date of birth. The postcode data element revealed a patient could have between two and 12 different postcodes within the datasets.

However, as the Quasi-ID was based on a concatenation of patient identifiers, Surname, Forename, Sex, Date of Birth and thus showed no variability within these details, whilst a patient may have between three and 12 postcodes over the course of seven years, which seems plausible.

Interestingly, the QID identified a unique patient might have between three and seven different Pseudo IDs, which indicates that within each dataset (and source CIS) the same patient was allocated multiple System Patient IDs, which may potentially result in an over-estimation of patient numbers if any analysis were based solely on unique Patient ID counts. Moreover, these findings provide further evidence to support the use of the Quasi-ID as the principal ID across the integrated dataset for this study.

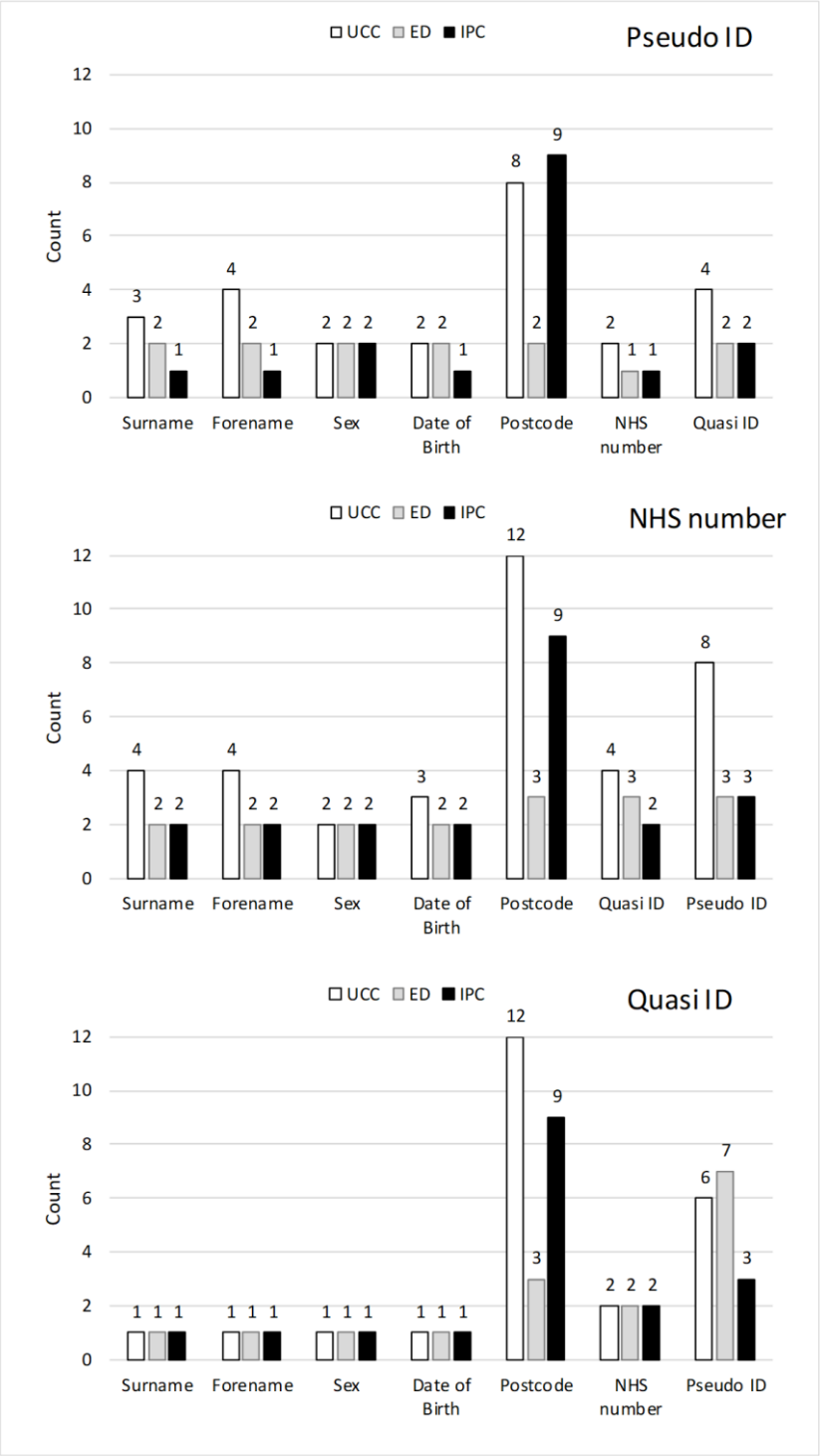


Figure 5.7 Variants of data elements per Pseudo ID, NHS number and Quasi-ID

5.6.1.3 Hybrid DRL deduplication

Individual patients were triangulated [335] within three patient identifiers, i.e. Pseudo ID, NHS number and Quasi-ID, and further interrelationships were identified between each of these IDs, adding a further level of complexity. For example, one patient may have two NHS numbers, three PIDs and one QID, resulting in duplication.

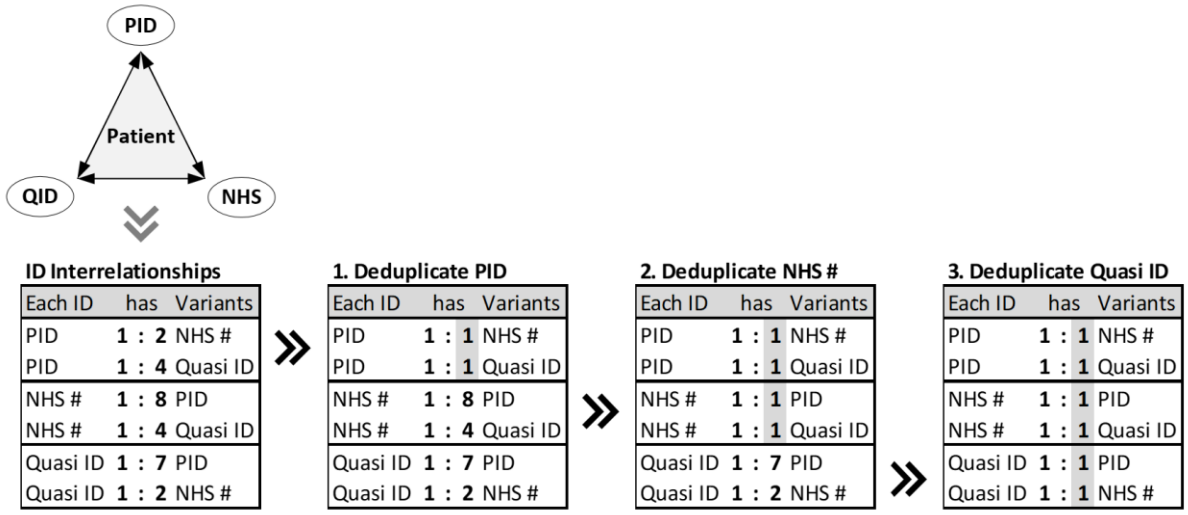


Figure 5.8 Deduplication of patient identification numbers.

Figure 5.8 (page 229) illustrates the process of deduplication of IDs by considering the perspective from each ID in relation to the remaining two IDs, e.g. consider deduplication via PID, whereby the ratio PID 1:2 NHS numbers, the frequency for each NHS number is calculated, and the modal NHS number (highest frequency of records) is promoted as the Master NHS number, resulting in the new ratio PID 1:1 NHS number. This process was repeated for the PID, NHS number and Quasi ID, respectively until all ID ratios were 1:1:1.

Table 5.16 (page 230) shows the number of unique patient records per each dataset post deduplication. The NHS number for the ED dataset recorded 32.8% completeness, providing further evidence of its inappropriateness for DRL only strategies.

Furthermore, deduplication via the ID identified some patients as having between zero and nine associated postcodes, i.e. one-to-many relationship. The repetition of patient details would inevitably result in a superficial increase in the number of unique record comparisons when using PRL strategies. This issue was addressed by transposing (reshaping from long to wide) the postcodes for each patient ID, thereby revealing the correct number of unique patients.

Following the completion of deduplication processes, the UCC dataset saw the greatest number of deduplicated records (8.9%), which reveals that unique patients were allocated more than one patient ID. Conversely, the IPC table saw the lowest number of deduplication (0.6%), which is as expected due to a dedicated team of people maintaining the parent IPC database.

A comparison of proportions between pre and post duplication, for each of the three datasets, revealed a significant reduction in the number of duplicate records within the UCC (P<0.0001), ED (P<0.0001) and IPC (P<0.05) datasets.

Table 5.16 Overview of unique record pre/post DRL deduplication.

	Dataset (N, %)			
	UCC	**ED**	**IPC**	**Total**
Total episodes (N)	**384213**	**498076**	**192792**	**1075081**
Unique patient records	256144 (66.67)	264692 (53.14)	114790 (59.54)	635626 (59.12)
Missing data	14	479	64	557
Unique records (excluding missing data)	256130 (66.66)	264213 (53.05)	114726 (59.51)	635069 (59.07)
Transposition of postcodes				
a. Unique records pre dedupe	242778 (63.19)	264130 (53.03)	108461 (56.26)	615369 (57.24)
Maximum postcodes per PID	8	2	9	--
b. Unique records post dedupe	**221080 (57.54)**	**259777 (52.16)**	**107847 (55.94)**	**588704 (54.76)**
Records with NHS number	171923 (70.81)	86574 (32.78)	99387 (91.63)	357884 (58.16)
Reduction a minus b	21698 (8.94)	4353 (1.65)	614 (0.57)	26665 (4.33)

Statistical analysis

Pre DRL deduplication	**UCC Table**	**ED Table**	**IPC Table**
Patients per 1000 episodes	632	530	563
Patients per 1 episode	0.632	0.530	0.563
95% Confidence Interval (lower, upper)	[0.630,0.633]	[0.529,0.532]	[0.560,0.565]
Post DRL deduplication			
Patients per 1000 episodes	575	522	559
Patients per 1 episode	0.575	0.522	0.559
95% Confidence Interval (lower, upper)	[0.574,0.577]	[0.520,0.523]	[0.557,0.562]
Z-Test of proportions (alpha = 0.05)	< 0.0001	< 0.0001	< 0.05

5.6.1.4 Hybrid PRL deduplication

Probabilistic linkage is based on deriving a match weight, which represents the likelihood of records belonging to the same individual, given the agreement or disagreement on a set of common identifiers. [204] The methods employed within this research use PRL techniques and referring to Figure 5.9 (page 231) and Figure 5.10 (page 232) show the distribution of match weights for Soundex Surname and Soundex Forename.

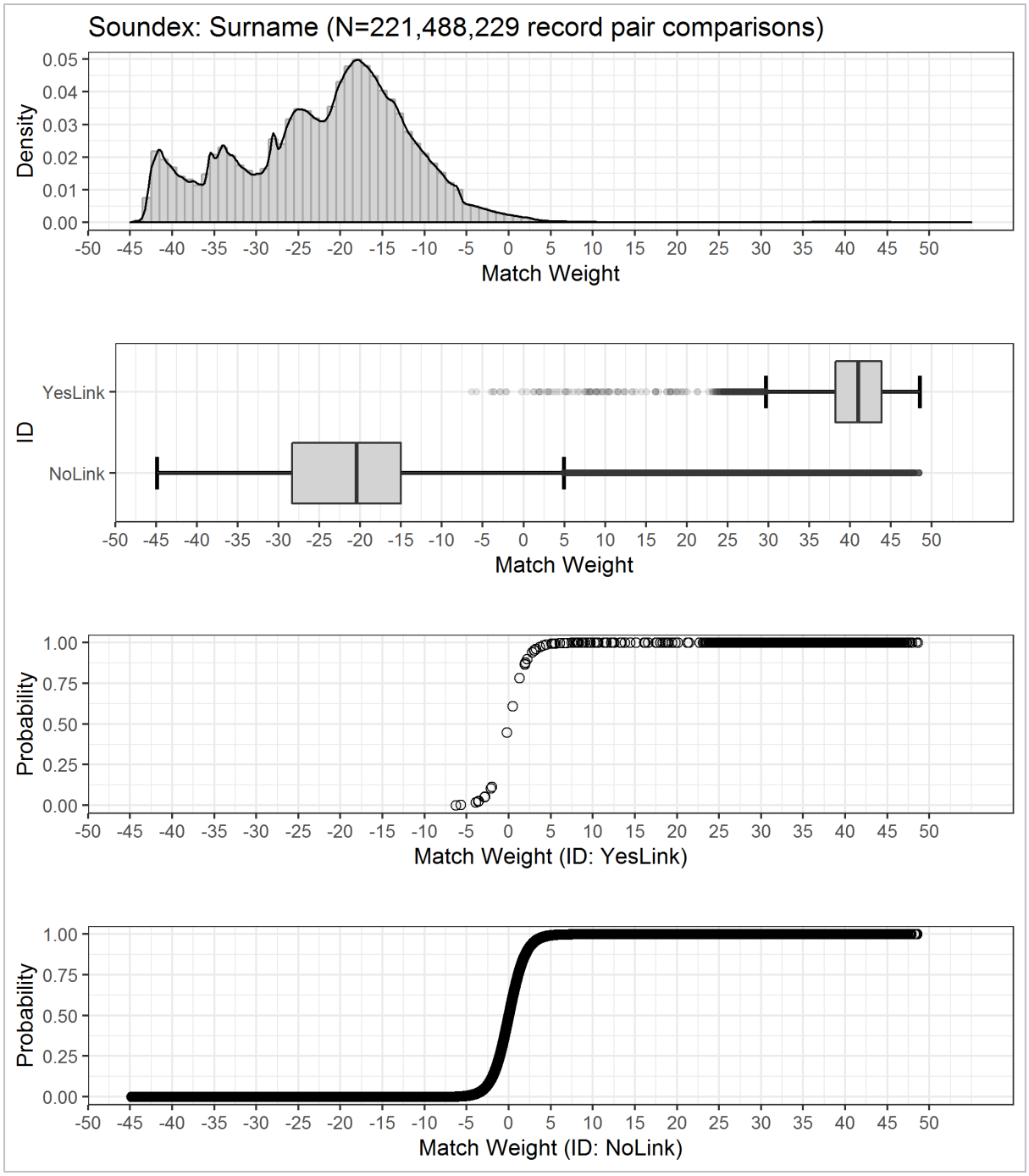


Figure 5.9 Distribution of match weights based on Soundex Surname

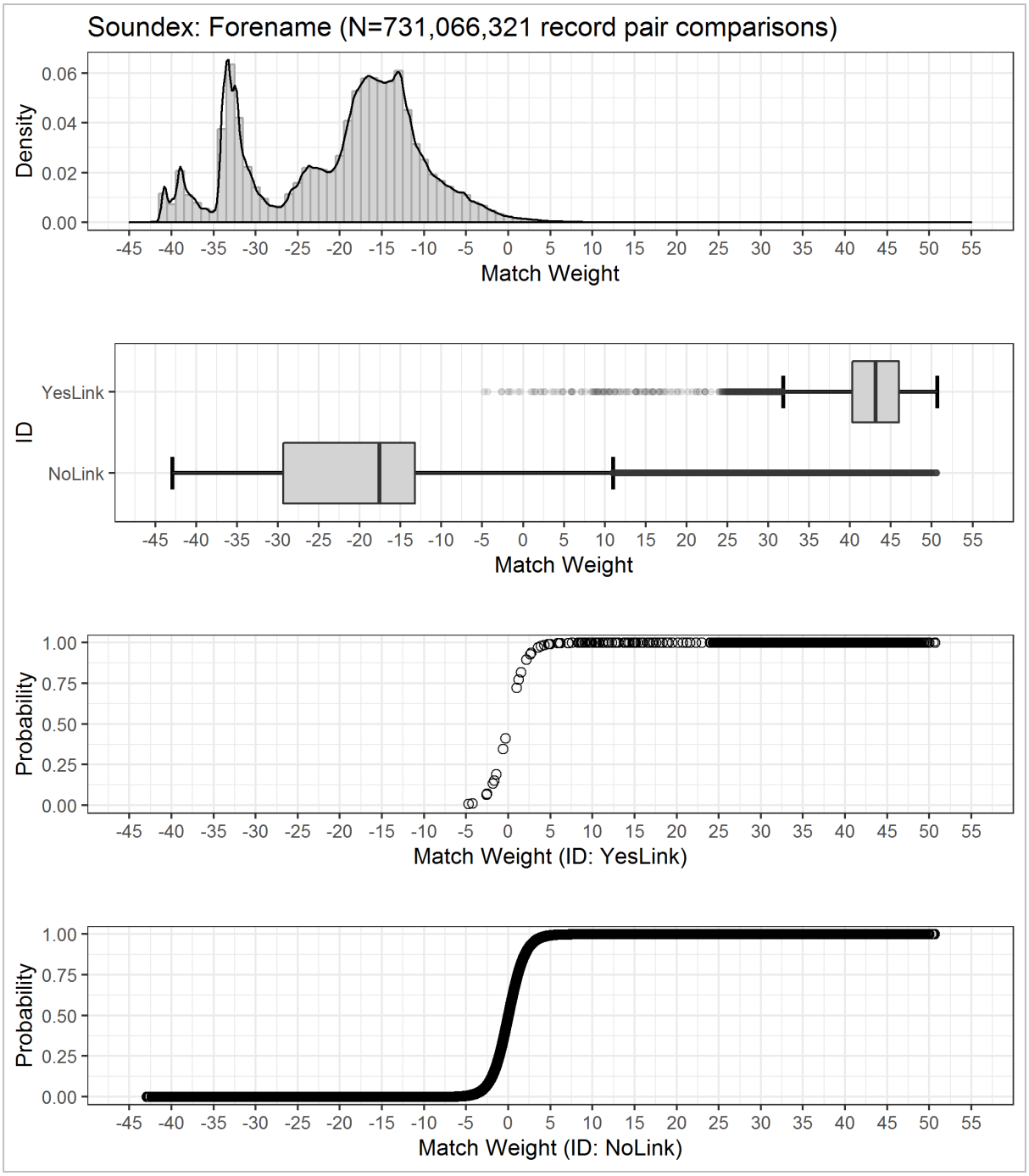


Figure 5.10 Distribution of match weights based on Soundex Forename

The density plots show the distribution of match weights, with the peaks showing locations of the highest concentration of points. The number of record pair comparisons with match weights less than five overshadow those with match weights greater than five.

Although the true match weights are hardly visible, the boxplot serves to provide an idea of the distribution, for example, forename depicts two distinct peaks and a wider variation as per the

interquartile range, when compared to surname. Furthermore, the box plot splits the match weights into two categories, in which the ID links or does not link. When ID is YesLink, there are outliers that range between -5 and 30; this indicates that although the ID linked, the demographic details may not always match. Hence, lower match scores may indicate that two different patients were inadvertently allocated the same ID, which would warrant further investigation. For clerical review, this study focused on record pairs whereby the ID did not match, however, the remaining data elements matched with a probability of 0.8 and above.

PRL process employs a Cartesian product; Soundex Surname resulted in approximately 221 million record pair comparisons, whereas Soundex Forename resulted in 731 million record pair comparisons. The purpose of PRL was to identify those records with the highest probability of being a match, i.e. identifying the same unique patient record, given that some demographic details may vary slightly due to typing errors, use of nicknames, the inclusion of maiden names or changes in surnames due to marriage.

Additionally, surnames typically have less variety, i.e. there are more chances of people with similar Surnames, whereas Forename is more unique and therefore contains far more variation, as reflected in the much higher number of comparisons.

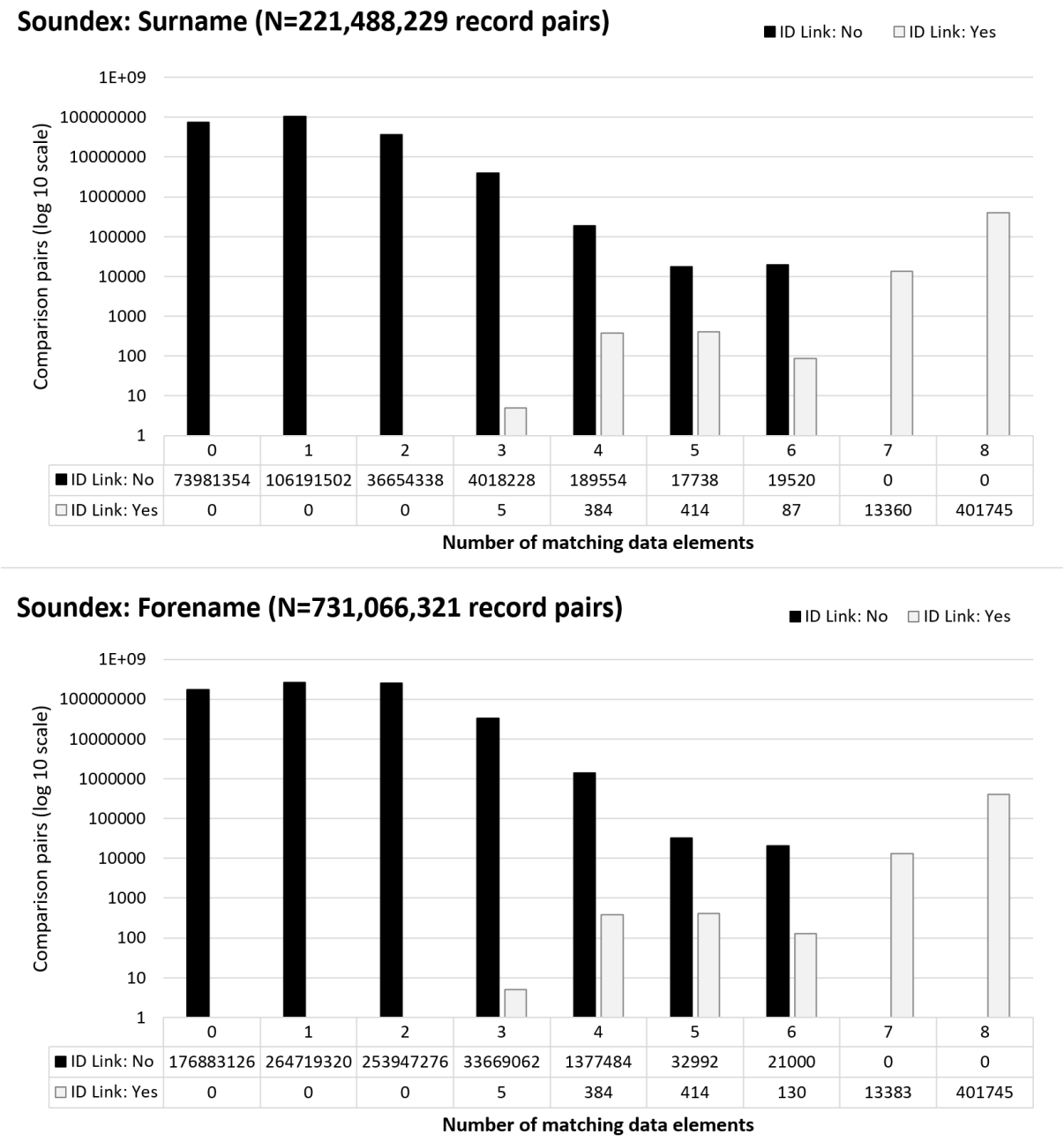
Referring to Figure 5.9 (page 231) and Figure 5.10 (page 232), the probability charts show a sigmoid shape between the match weights ranging from -5 (probability < 0.1) through 0 (probability 0.5) to +5 (probability > 0.8).

Therefore selecting those record pair comparisons where category ID is NoLink, the higher match weights and correspondingly high probability strongly indicate the demographics in both records belong to the same individual, who has inadvertently allocated more than one ID.

Thus, adopting a conservative approach, record pairs with a match probability of ≥ 0.8 were identified/selected from the results of both Soundex Surname and Soundex Forename comparison pairs for clerical review, to facilitate the deduplication process.

Figure 5.11 (page 234) illustrates the number of comparison pairs per matching data elements. The black bars indicate records where the IDs do not link, i.e. they are the result of the Cartesian product showing all possible combination pairs.

For example, focusing on Soundex Surname, the black bars exclude the ID, and represent the remaining seven data elements that can potentially match, at least six of these (N=19520) have resulted in a match, which is a good indicator of potential missed matches that require clerical review. Thus, for this study only comparison pairs (black bars) with four or more matching data elements were selected for further investigation and clerical review.



	0	1	2	3	4	5	6	7	8
ID Link: No	73981354	106191502	36654338	4018228	189554	17738	19520	0	0
ID Link: Yes	0	0	0	5	384	414	87	13360	401745

	0	1	2	3	4	5	6	7	8
ID Link: No	176883126	264719320	253947276	33669062	1377484	32992	21000	0	0
ID Link: Yes	0	0	0	5	384	414	130	13383	401745

Figure 5.11 Soundex Surname & Soundex Forename matching data elements

5.6.2 Comparison of linkage strategies

Following the application of the HES and the Hybrid algorithm, Figure 5.12 (page 235) illustrates the effect of both algorithms at linking patient records across each of the three datasets and compares the results relative to the number of correctly identified records belonging to the same individual across multiple datasets.

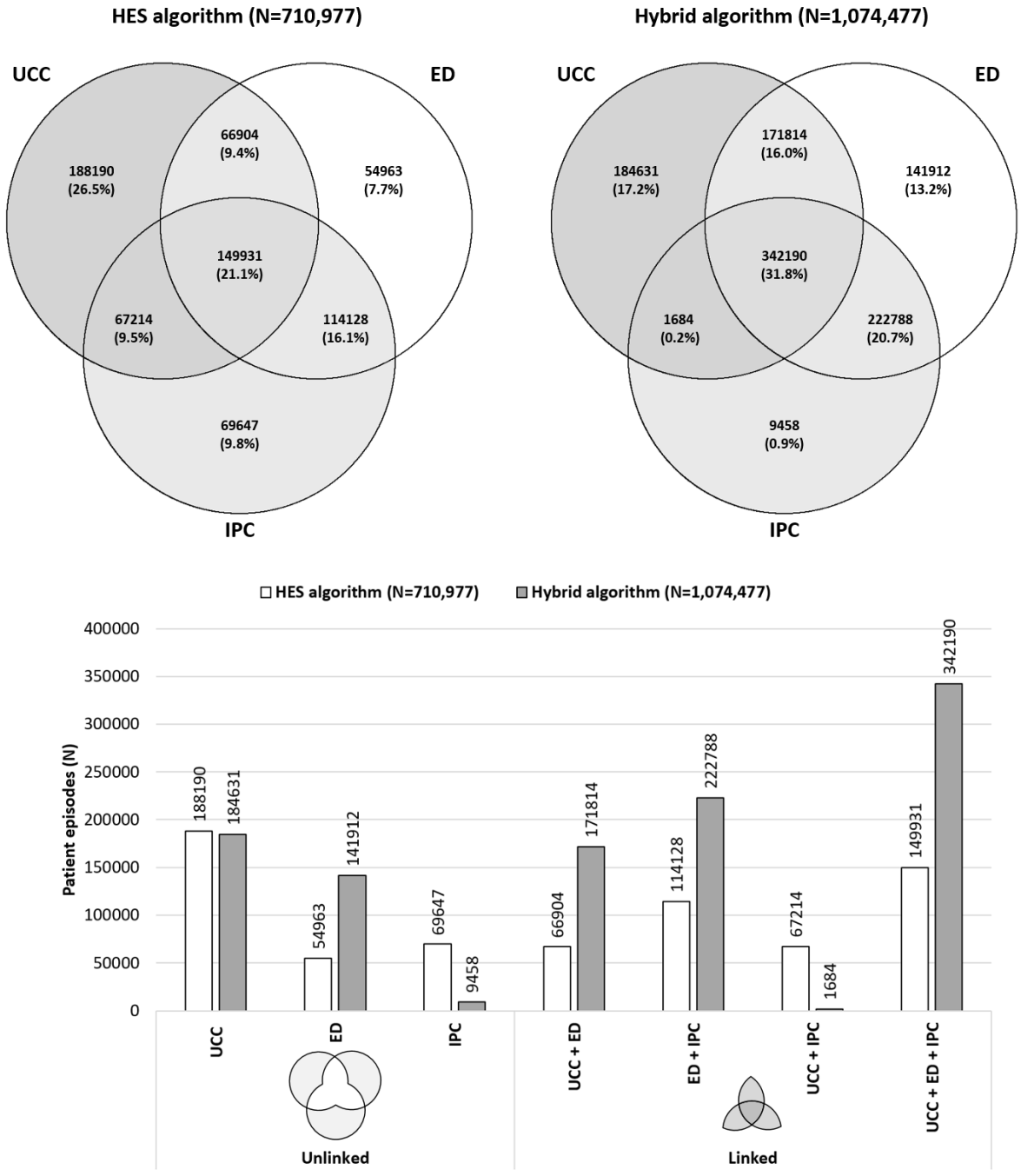


Figure 5.12 Comparison of HES algorithm with the Hybrid algorithm

The HES algorithm linked approximately 711 thousand records, whereas the Hybrid algorithm was able to link 1.07 million records. A distinction is made between Linked which refers to areas in which two or three datasets intersect and overlap, e.g. HES algorithm UCC and ED intersection, n=66904, and Unlinked refers to the segments which do not overlap with any other dataset, i.e. records remain exclusively within one dataset, e.g. HES algorithm UCC is n=188190.

When comparing the central intersection of all three datasets, the Hybrid algorithm resulted in a higher proportion of matches 31.8%, which is comparatively higher than the HES algorithm, which matched only 21.1% of records. The bar chart further serves to illustrate the effect of each algorithm on the linkage of the three datasets.

Objective six of this thesis was to differentiate empirically whether the selection and application of a particular linkage strategy influenced the number of correctly identified records belonging to the same individual across multiple datasets. Thus, the following hypothesis (page 21) was tested using the Two Proportion Z-Test, which is a type of hypothesis test to compare proportions from two independent samples to see if they are the same. [179]

H_0: *There is no difference between using deterministic or probabilistic record linkage techniques.*
H_1: *There is a difference between using deterministic or probabilistic record linkage techniques*

Table 5.17 Comparison of linkage approaches

	Departments							
	Unlinked			**Linked**				
	UCC	**ED**	**IPC**	**UCC +ED**	**ED +IPC**	**UCC +IPC**	**UCC+ED +IPC**	**Total**
HES Algorithm								
(a) Episodes	188,190	54,963	69,647	66,904	114,128	67,214	149,931	710,977
(b) Unique patients	130,481	32,920	50,234	15,433	23,050	16,637	14,831	283,586
Rate: (b/a)*1000	**693**	**599**	**721**	**231**	**202**	**248**	**99**	**399**
95% Confidence Interval	[0.691, 0.695]	[0.595, 0.603]	[0.718, 0.725]	[0.227, 0.234]	[0.200, 0.204]	[0.244, 0.251]	[0.097, 0.100]	[0.398, 0.400]
Hybrid approach								
(c) Episodes	184,631	141,912	9,458	171,814	222,788	1,684	342,190	1,074,477
(d) Unique patients	134,665	114,384	8,357	46,529	59,605	636	39,249	403,425
Rate: (d/c)*1000	**729**	**806**	**884**	**271**	**268**	**378**	**115**	**375**
95% Confidence Interval	[0.727, 0.731]	[0.804, 0.808]	[0.877, 0.890]	[0.269, 0.273]	[0.266, 0.269]	[0.355, 0.401]	[0.114, 0.116]	[0.375, 0.376]
Z-Test of proportions (alpha = 0.05)								
P value	****	****	****	****	****	****	****	****

***** = p<0.0001*

Table 5.17 (page 236), provides a comparison of the ability of the two algorithms at identifying the number of unique patients per total episodes expressed as a rate per 1000, for all intersections of the three datasets. The results were compared using the Two Proportion Z-Test.

Therefore, comparing each intersection, within individual datasets and across multiple datasets, the z-score associated with a 5% alpha level was found to be statistically significant ($P<0.001$) for each category. Thus, disproving the null hypothesis and concluding that there is a difference between the using deterministic (HES algorithm) or probabilistic (Hybrid algorithm) record linkage techniques, in correctly identifying patients across all three datasets.

Moreover, this difference may also be attributed to the fact that as the HES algorithm specifically requires the presence of the NHS number to undertake data linkage, and missing NHS numbers will inadvertently affect results. However, the Hybrid algorithm is not dependent upon the NHS number and was designed to use the Quasi-ID, created from patient demographic details commonly found across all three datasets, hence, it was able to yield and identify a significantly higher number of linkages and unique individuals compared to the HES algorithm.

5.6.3 Linked dataset

The selection and application of the Hybrid algorithm resulted in the creation of one Master Patient Table that contained records for every individual patient across all three UCC, ED and IPC datasets, uniquely identified via the Quasi-ID. The Master Patient Table facilitated the connection of the disparate datasets to create one Linked dataset, which did not contain any duplicate records, thus, addressing the issue of double-counting and overestimation of patient numbers.

An example of the Linked dataset record structure is provided in Figure 5.13 (page 238), in which records are patient-oriented and structured in a similar design to the 'Electronic Health Record' providing a longitudinal record of a patient's health. Each patient can be identified uniquely via the Quasi-ID which is in turn, linked to each Patient Episode and subsequent Clinical Encounter [344] across the UCC, ED and IPC departments.

Furthermore, each Encounter places emphasis upon the date/time of attendance and/or admission, as recorded within each disparate CIS, permitting the arrangement of the episodes and encounters

in chronological order. From a database perspective, a one-to-many relationship was established, whereby one patient was associated with many episodes/encounters of care.

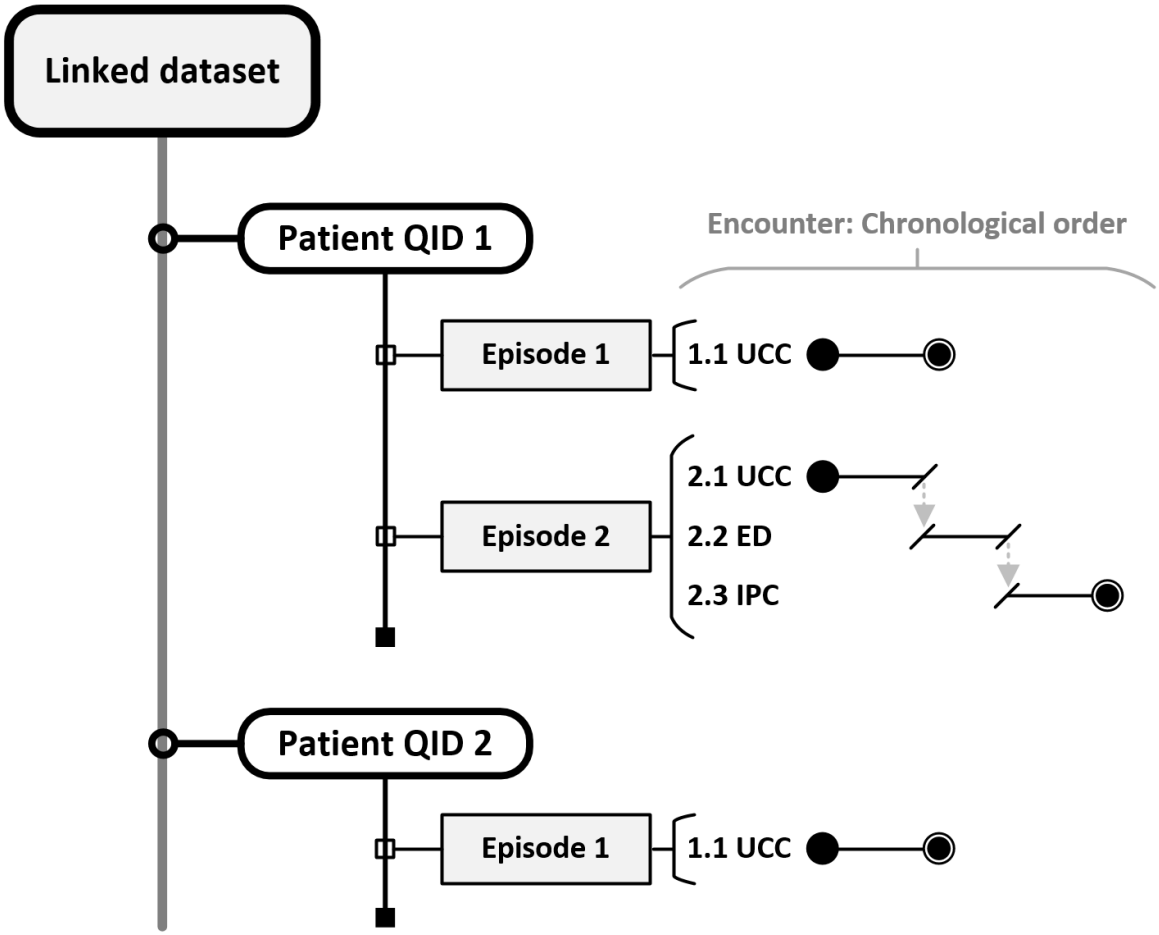


Figure 5.13 Structure of patient records and associated episodes records within the Linked dataset.

For example, 'Patient Quasi-ID 1' has two distinct Episodes of care, and Episode 2 comprises three encounters, Encounter 2.1, the patient presented to the UCC and was referred to ED, Encounter 2.2, the patient was treated in ED but required admission, and Encounter 2.3, the patient was admitted to IPC and eventually discharged home.

Moreover, the data element Start date/time is common to all three datasets, and sorting the dates/time in ascending order, i.e. from the earliest to latest record, permits the chronological arrangement of episodes/encounters from the three disparate datasets. Additionally, to demonstrate the practical applicability of the Linked dataset, data analyses as per Objectives 7, 8 and 9 of this thesis were undertaken and presented in the following sections.

5.7 Linked dataset analysis

To demonstrate the practical applicability of the Linked longitudinal dataset, three distinct analyses were undertaken, and results presented henceforth.

5.7.1 Transition times

Addresses objective 7: To calculate the transition time for patient transfer between urgent and emergency care services

Within the Linked dataset, patient episodes and encounters were arranged in chronological order, which enabled the identification of a previously unmeasurable phenomenon termed the 'Transition Time'.

The UCC and ED are co-located at both CXH and HH sites and share the same reception. For example, once a patient has been seen in the UCC and needs referral to ED, the patient must physically walk back to the main reception, where the ED receptionist will register the patient's details into the ED database. The time taken for this process is the 'Transition time' from one service to the next, and a simple algorithm was created/applied to the Linked dataset to check for instances of cases in which the departure time from one system was greater than the arrival time within the subsequent system.

5.7.1.1 Transition time calculation

Table 5.18 Example of Transition time calculation

			Time (minutes)	
Department	**Arrival Date/Time**	**Depart Date/Time**	**Time in department**	**Transition time**
UCC	14/02/2011 10:18	14/02/2011 **10:22**	4	
ED	14/02/2011 **10:39**	14/02/2011 11:55	76	**17**

Calculation of the 'Transition Time' as per Table 5.18 (page 239), it took one patient an extra 17 minutes from being departed from the UCC service, to queue up, speak to an ED receptionist, re-register and commence a new episode of care within the ED database, prior to receiving treatment within ED.

5.7.1.2 Transition time from UCC to ED

Figure 5.14 (page 240) depicts a boxplot comparing the transition times between UCC patient referral to ED, by hospital site and reveals how long a patient must wait from UCC referral to registration at ED reception, and having their details transcribed within the ED database.

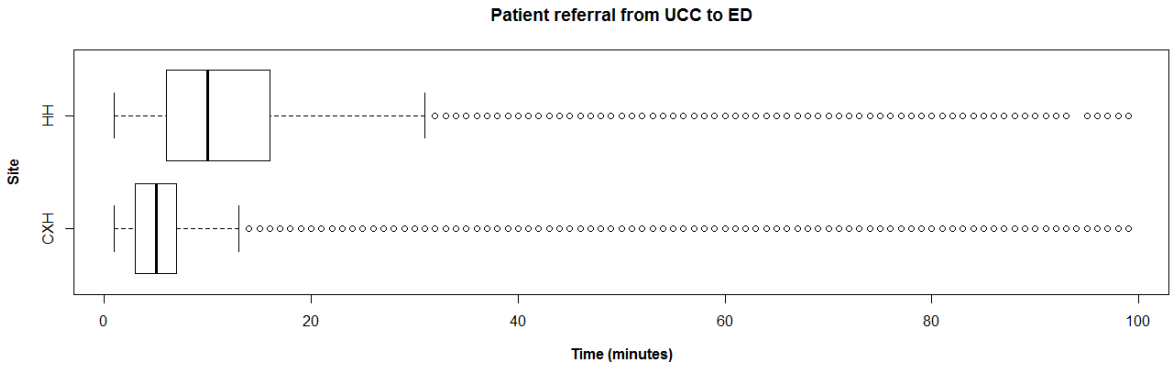


Figure 5.14 Boxplot of transition times from UCC referral to ED across CXH and HH sites.

The results from the independent group t-test were calculated, and the mean, 95% Confidence Intervals (CI) and Median values summarised as follows:

- CXH: Mean 6.99 minutes, (95% CI: 6.90, 7.09), Median 5 minutes.
- HH: Mean 13.40 minutes, (95% CI: 13.20, 13.60), Median 10 minutes.

These results show a statistically significant difference (P<0.0001) in the mean transition time between CXH and HH sites. Additionally, the mean time at HH is almost double that of CXH, similarly reflected in the 50th Percentile, whereby CXH is 5 minutes, and HH is 10 minutes.

This difference may be attributable to the physical sizes of each respective hospital site and department, CXH is comparatively larger than HH and has more reception desks; thus, patients wait for less time to be seen, hence the shorter transition times. However, there a large number of outliers which may be reflective of busy periods in the department or perhaps cases in which patients decided to return to ED at a later time, or travelled cross-site.

5.7.1.3 Referral time from ED to IPC

The boxplot illustrated in Figure 5.15 (page 241) compares the time between referrals from ED to IPC for inpatient admission, by hospital site, and reflects how long a patient must wait for a bed to become available, transported to IPC and have their details transcribed within the IPC database.

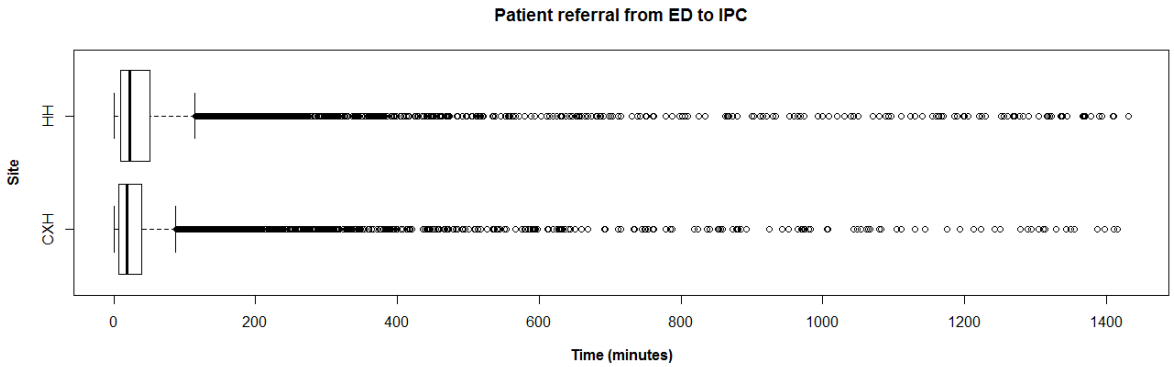


Figure 5.15 Boxplot of referral times from ED to IPC across CXH and HH sites.

The results from the independent group t-test were calculated, and the mean, 95% CI and Third Quartile (P75) values summarised as follows:

- CXH: Mean 44.12 minutes, (95% CI: 42.07, 46.16), P75: 39 minutes.
- HH: Mean 55.64 minutes, (95% CI: 53.21, 58.08), P75: 51 minutes.

These results show a statistically significant difference ($P<0.0001$) in the mean referral time between CXH and HH sites. In addition, the mean time at HH is higher than CXH by approximately 11.53 minutes, reflected correspondingly in the difference between the 75th Percentiles at both sites.

However, there a large number of outliers for both sites, which may be indicative of particularly busy periods within the hospital, and reflect the time required for availability and preparation of beds within IPC, or certain patient conditions that necessitate referral to speciality admissions.

Additionally, each hospital site contains specialties, e.g. Renal or Cardiology units at HH site, so patients may require transport cross-site before they are admitted to IPC. Additional indirect factors that may affect referral time include the availability of specialist staff, hospital size/layout, day of the week, and the number of supporting staff, e.g. porters.

5.7.1.4 Referral time from ED to UCC

The boxplot shown in Figure 5.16 (page 242) compares, by hospital site, the time between referrals from ED to UCC for patients initially arriving by ambulance, triaged within ED and following assessment, their condition identified as more appropriate for treatment within UCC services and subsequently having their details transcribed within the UCC database.

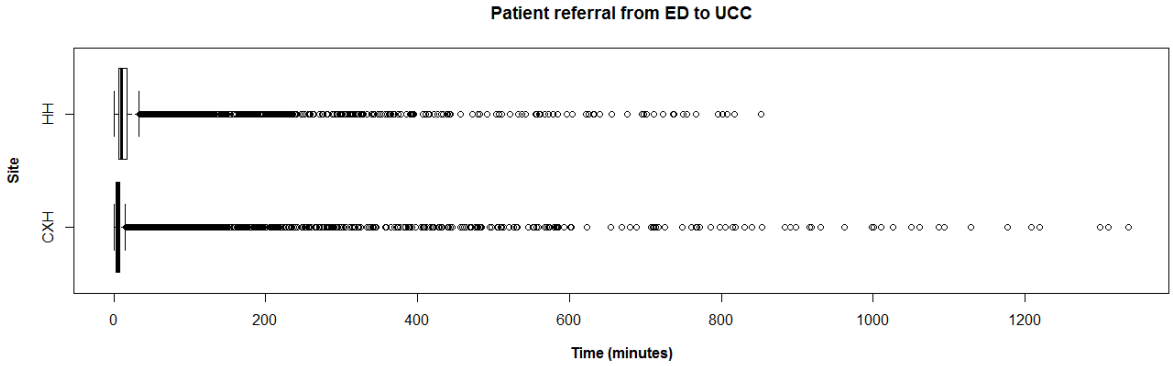


Figure 5.16 Boxplot of referral times from ED to UCC, across CXH and HH sites.

The results from the independent group t-test were calculated, and the mean, 95% CI and Third Quartile (P75) values summarised as follows:

- CXH: Mean 11.05 minutes, (95% CI: 10.54, 11.55), P75: 8 minutes.
- HH: Mean 19.64 minutes, (95% CI: 18.89, 20.40), P75: 17 minutes.

These results show a statistically significant difference (P<0.0001) in the mean referral time between CXH and HH sites. Additionally, the mean time at HH is higher than CXH by approximately 8.60 minutes, also reflected in the difference between the 75th Percentiles at both sites.

Thus, 75% of patients are able to see a UCC receptionist and have their details transcribed within 17 minutes at the HH site and 8 minutes at CXH site. Again, the difference in times may be attributable to the size of each respective department, with CXH UCC being comparatively larger and equipped with more reception desks than HH, so patients wait for less time.

Interestingly, the interquartile range for CXH is between 3 to 8 minutes, and for HH is between 6 to 17 minutes, both boxes are fairly narrow, however, there a large number of outliers at each site which may reflect busy periods of time.

5.7.2 ED four-hour standard

Addresses objective 8: To distinguish the effect of utilising Linked and Unlinked datasets upon the ED 4 hour standard.

5.7.2.1 Chronological discrepancies due to Date/time recording

Each individual CIS within UCC, ED and IPC services captures the Arrival date/time and Depart date/time associated with each patient attendance or admission. However, as the Linked dataset is designed from the patient perspective, it does not rely on the linkage of specific dates/times, rather it chronologically aligns all patient records from across three services to provide a longitudinal episode of care. This distinctive feature enables the identification of chronological discrepancies between the dates/times for patients traversing more than one service.

Table 5.19 Chronological discrepancies between services.

			Time (minutes)	
Department	**Arrival Date/Time**	**Depart Date/Time**	**Time in department**	**Corrected Time**
ED	04/03/2010 07:50	04/03/2010 **17:03**	553	**146**
IPC	04/03/2010 **10:16**	04/03/2010 16:57	401	**401**

For example, Table 5.19 (page 243) presents the total length of time a patient spent in ED from Arrival to Depart, i.e. 553 minutes, which is clearly over four hours and would be classified as a 'Breach' of the 95% target. The process for Breach validation requires an analyst to identify the breach and refer the specific case to a senior Clinician, who would spend further time reviewing case details in order to confirm that the case was, in fact, a valid breach.

However, using the Linked dataset, it is evident that the patient had been admitted to hospital, and their episode of admission within the IPC database commenced just over two hours after being seen within ED. Hence, the correct time the patient departed ED would have been just after 10:00 am, which would not have flagged up as a breach, thus, negating any requirement for the senior Clinician to perform any breach investigations whilst saving valuable time.

Furthermore, the Linked dataset had successfully identified many chronological discrepancies due to patient referrals between services. Therefore, to investigate the effect of reconciling these discrepancies, the 'Total time spent in ED' and Four-hour standard (Median, 95th Percentile, single

longest time) were calculated per site, utilising the CDW Linked dataset and an Unlinked ED dataset, to enable comparison of these metrics.

5.7.2.2 HH ED effect of chronological reconciliation

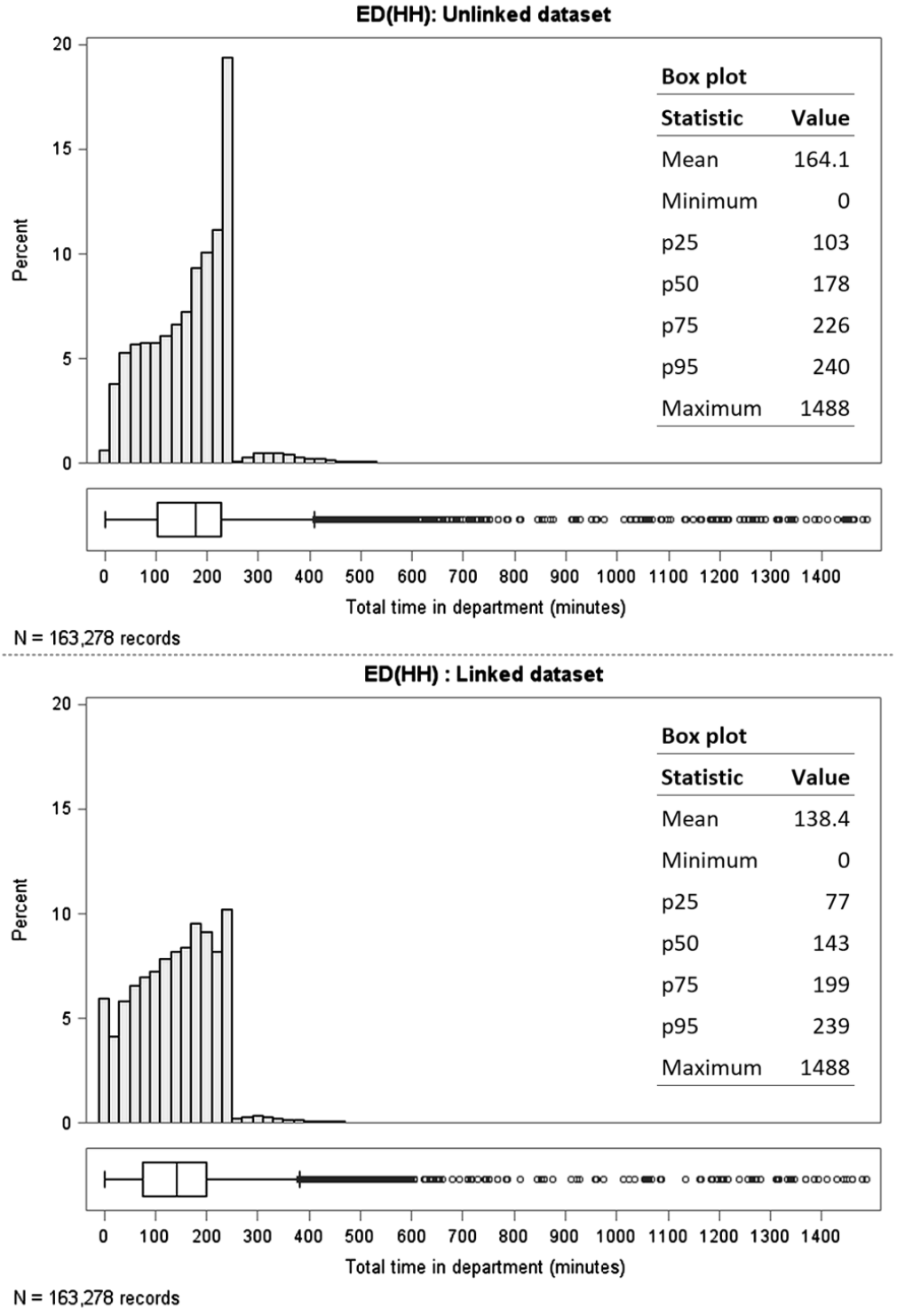


Figure 5.17 HH ED comparison of time in ED within Linked and Unlinked datasets

5.7.2.3 CXH ED effect of chronological reconciliation

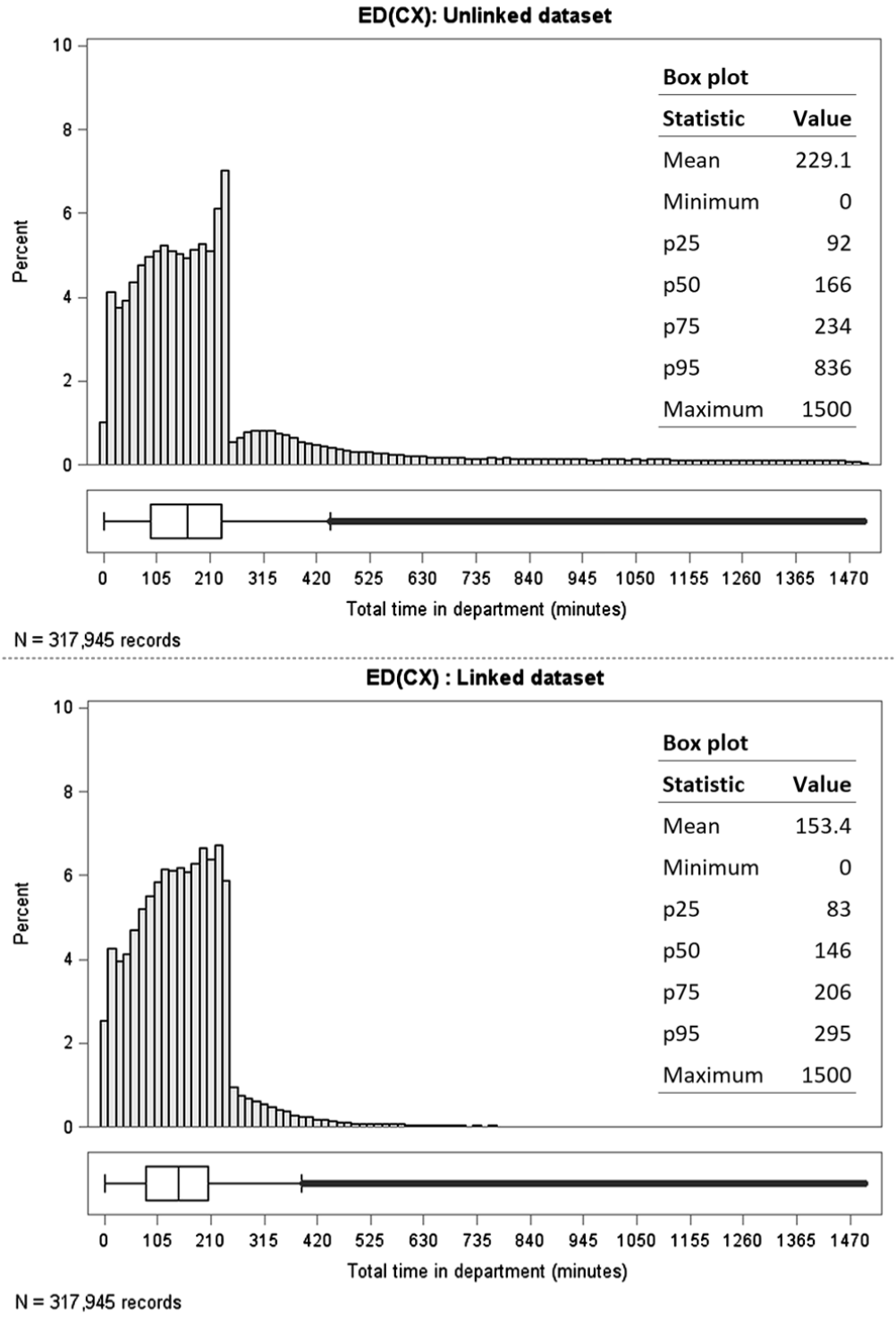


Figure 5.18 CXH ED comparison of time in ED within Linked and Unlinked datasets

Table 5.20 Comparison of results from Linked and Unlinked datasets

Metric	Unlinked dataset	Linked dataset
HH Site		
>> Total time in ED	Mean 164.14 minutes (95% CI: 163.74, 164.54)	Mean 138.40 minutes (95% CI: 138.01, 138.79)
	t-test independent groups: significant difference (P<0.0001)	
Four-hour standard		
>> Median, 50th Percentile	178 minutes	143 minutes
>> 95th Percentile	240 minutes	239 minutes
>> Single longest time	1488 minutes	1488 minutes
CXH Site		
>> Total time in ED	Mean 229.14 minutes (95% CI: 228.26, 230.03)	Mean 153.42 minutes (95% CI: 153.03, 153.82)
	t-test independent groups: significant difference (P<0.0001)	
Four-hour standard		
>> Median, 50th Percentile	166 minutes	146 minutes
>> 95th Percentile	836 minutes	295 minutes
>> Single longest time	1500 minutes	1500 minutes

5.7.2.4 HH site summary

Referring to Figure 5.17 (page 244) the unlinked dataset shows a high proportion of patients, almost 19% of all records, before the four-hour (240 minutes) marker, which drops suddenly to near 1% after 241 minutes. This is because staff within EDs continuously monitor the 4-hour standard and aim to complete cases where possible for patients nearing that limit, in order to avoid potential breaches. However, when looking at the chart for the Linked dataset, the most noticeable difference is the drop to almost 10% of all records near 240 minutes.

Furthermore, Table 5.20 (page 246) presents the results from the comparison of total time in ED as calculated using the Unlinked and Linked datasets, respectively. These results show a statistically significant difference (P<0.0001) between the two datasets. The Linked dataset reduced the mean time in department by 25.7 minutes, and the Median by 35 minutes, although the effect on the 95th Percentile was marginal. Thus, demonstrating the effect of chronological reconciliation using the Linked dataset.

5.7.2.5 CXH site summary

Referring to Figure 5.18 (page 245) the unlinked dataset, shows an almost bimodal distribution, the first distribution has a high number of records (almost 8%) pre 240 minutes, and the second distribution post 240 minutes is comparatively more flat peaking near 1% but has a wider range that remains fairly steady reaching 1470 minutes. Compared to the Linked dataset, the effect on the range within the second distribution is clearly visible, with the percentage of records after 240 minutes decreasing by almost half in range to approximately 700 minutes. Furthermore, looking at the percentage of patients near the 240-minute mark, the percentage of records has fallen from 7% in the Unlinked dataset, to approximately 5.7% in the Linked dataset.

Moreover, the results from Table 5.20 (page 246) show a statistically significant difference ($P<0.0001$) between the two datasets. The Linked dataset reduced the mean time in department by 75.75 minutes, and the Median by 20 minutes, although the largest effect was on the 95th Percentile, which fell from 836 to 295 minutes, a reduction of 541 minutes. Thus, again demonstrating the effect of chronological reconciliation using the Linked dataset.

5.7.2.6 The benefit of the Linked dataset

The metrics, total time in ED and the four-hour standard were used as a proxy to observe the effect of utilising the Linked dataset versus the unlinked dataset as a comparator. However, the Linked dataset contained a conditional statement (detailed on page 197) which automatically identified and reconciled all chronological discrepancies, resulting in more accurate ED metrics, which were significantly lower when compared to the same results from the Unlinked dataset.

Thus, using the Linked dataset provides a genuine reflection of the four-hour metrics, whilst reducing the number of potential four-hour breaches that require clinical review, saving valuable clinical time.

Nonetheless, this example clearly demonstrates the potential benefits associated with using the Linked dataset for analysis purposes and serves to provide an element of validating the quality of the Linked dataset to ensure its accuracy in measuring episode Start/End date/times, and provides assurance that results from any subsequent analysis are trustworthy.

5.7.3 Comparison of admission rate pre/post UCC

Addresses objective 9: To compare the rate of emergency patient admissions pre/post UCC implementation

Within this section, the Linked dataset was utilised to investigate the rate of emergency patient admissions pre/post UCC implementation, using an Interrupted Time Series (ITS) model for each site.

5.7.3.1 HH site

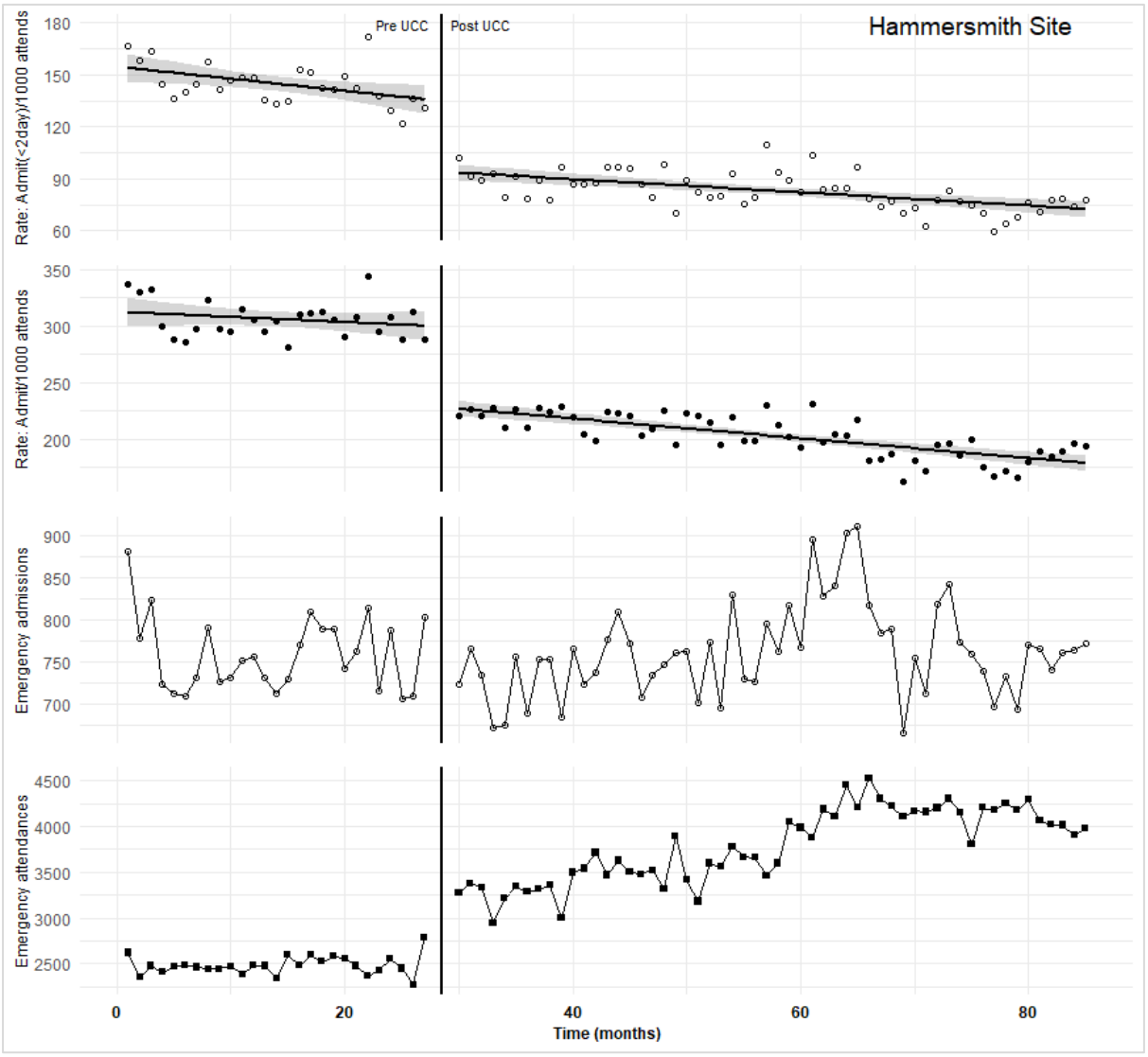


Figure 5.19 Episodes of attendance and admissions pre/post UCC launch at HH site

Figure 5.19 (page 248) provides an overview of the number of episodes of attendance and admission, along with the total admission rate and short-stay admission rate, per month throughout the study period, pre/post UCC service implementation at the HH site, the post-transition month

was not included. Interestingly, post-UCC, the number of ED attendances increase steadily, whist emergency admissions remain fairly consistent. However, post-intervention, the overall admission rate and short-stay admission rate (< 48 hours) both show a step change and downward slope trend.

Referring to Table 5.21 (page 252) during the study period, there were a total of 278,783 urgent and ED attendances and 63,127 emergency hospital admissions. Short stay admissions accounted for 43.0% of all emergency admissions.

There were a higher number of attendances by female (57.0%) patients, and the most common patient age group was 25 to 39 year age group (34.2%), reflective of the local population in H&F borough. Children aged between 0 to 14 years had the highest ratio of attendance to admission (188 attendances per one admission), in contrast with those aged 80+ years, who had the lowest ratio, i.e. every 1.6 attendances resulted in an admission.

The mean monthly number of total ED attendances was 2,482 pre-UCC rising to 3,782 attendances post-UCC; an overall increase of 52.4%. In contrast, mean emergency admissions rose by 0.3% from 759 to 761 per month. Pre-UCC, 30.6% of ED attendances resulted in an admission (N=67,012), post-UCC this figure fell to 20.1% of attendances (N=211,771) requiring hospital admission.

Table 5.22 (page 253) shows that pre-UCC, there were no significant yearly changes in the number of ED attendances, i.e. 0.65 attendances per year (95% CI: -1.70, 3.00). After the implementation of the UCC, there was a significant level change in the number of attendances (regression coefficient 11.80 attendances per unit of analysis; 95% CI: 8.00, 15.60) and a change in time trend (regression coefficient 3.5 attendances per unit of analysis; 95% CI: 1.03, 5.97).

These results were equivalent to an 18.1% increase in the number of attendances within the first year of UCC implementation, and another 5.4% increase each year at the HH site, based on the average number of attendance over the study period which was 65 attendances per unit of analysis at HH.

Referring to Table 5.23 (page 253) and Figure 5.19 (page 248) which show that pre-UCC there was a trend in emergency hospital admissions falling; a reduction of 5.0% with adjusted IRR 0.95 (95% CI: 0.93, 0.97). Post-UCC, there was an 18.0% reduction (level change) in the overall hospital admissions rate (adjusted IRR 0.82, 95% CI: 0.79, 0.85). This comprised of a 27.0% drop in the short stay

admission rate (adjusted IRR 0.73, 95% CI: 0.70, 0.77) and 10.0% in long stay admission rate (adjusted IRR 0.90, 95% CI: 0.85, 0.94) without any significant changes in slope.

5.7.3.2 CXH site

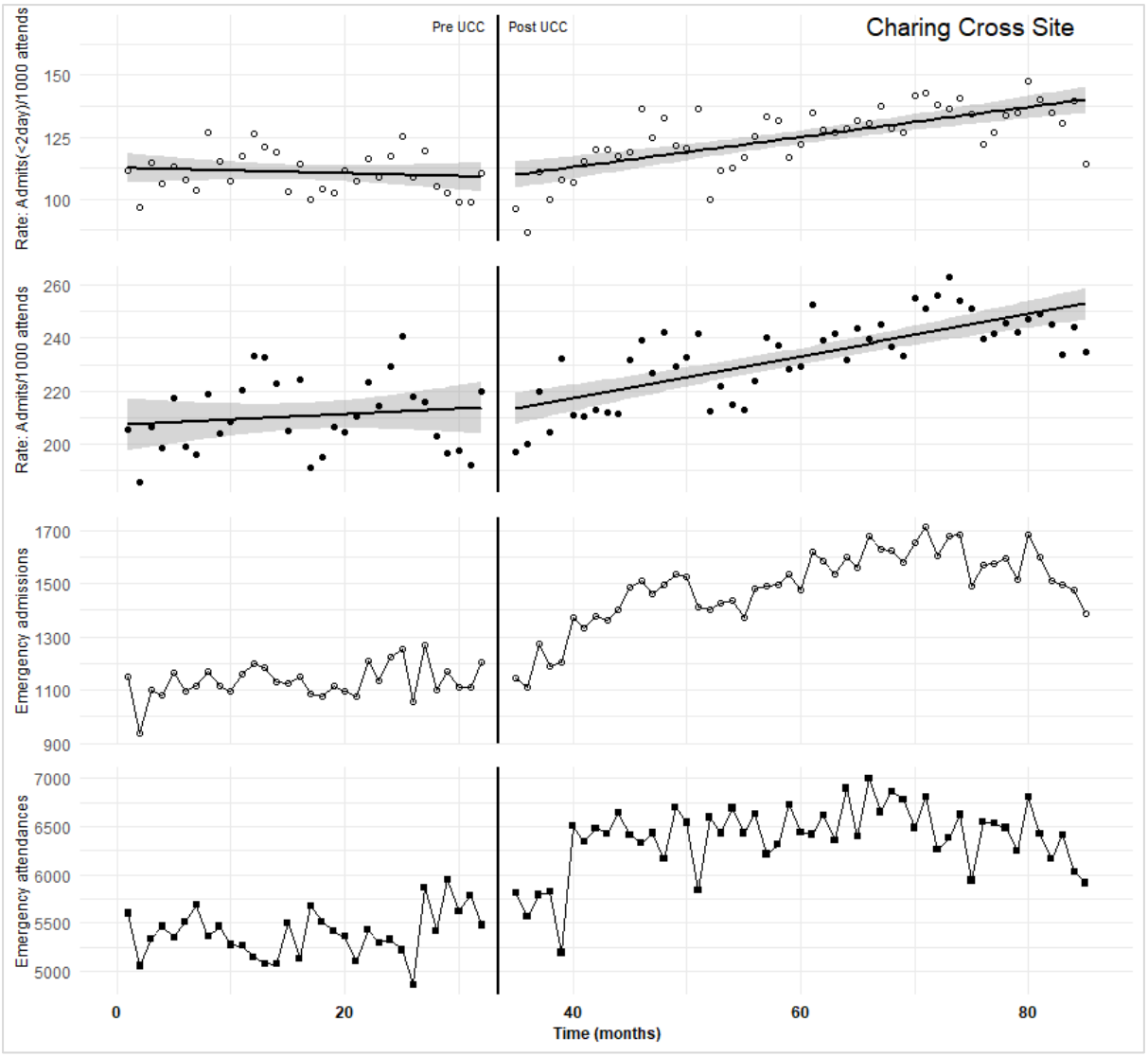


Figure 5.20 Episodes of attendance and admissions pre/post UCC launch at CXH site

Figure 5.20 (page 250) provides an overview of the number of episodes of attendance and admission, along with the total admission rate and short-stay admission rate, per month throughout the study period, pre/post UCC service implementation at the CXH site, the post-transition month was not included.

Post-UCC, the number of attendances and admissions, increase up to month 80 after which they appear to stabilise. However, both the overall and short-stay admission rates show an upward slope trend with no discernible step change, which is in contrast to the HH site.

Referring to Table 5.21 (page 252), there were 498,144 urgent and ED attendances and 111,575 emergency hospital admissions. Short stay admissions accounted for 53.0% of all emergency admissions.

Attendances were similar among female (51.0%) and male patients, with the majority of patients falling within the 25 to 39 year-old age group (34.8%). Additionally, Children aged between 0 to 14 years had the highest ratio of attendance to admission (1,573 attendances per one admission) compared with those aged 80+ years, who had the lowest ratio, i.e. every 1.4 attendances resulted in an admission.

The mean monthly number of total ED attendances rose from 5,395 pre-UCC to 6,382 attendances post-UCC; an increase of 18.3%. In addition, emergency admissions rose by 30.2% from 1134 to 1476 per month. Pre-UCC, 21.0% of ED attendances were admitted (N=172,648), post-UCC this figure rose to 23.1% of attendances (N=325,496), requiring hospital admission.

Table 5.22 (page 253) shows that pre-UCC, there were no significant yearly changes in ED attendances; 0.78 attendances per year, (95% CI: -1.07, 2.63). Post-UCC, the level change showed a significant increase of 12.77 attendances per year (95% CI: 9.06, 16.48). These were not associated with changes in underlying time trend; 0.59 attendances per year (95% CI:-1.46, 2.64). These results were equivalent to a 12.2 % increase in the number of attendances in the first year of UCC implementation, based on the average number of attendance being 105 attendances per unit of analysis at CXH site.

Referring to Table 5.23 (page 253) and Figure 5.20 (page 250) which shows pre-UCC that there was no significant time trend in hospital admission rates prior to the implementation of the UCC (adjusted IRR 1.00; 95% CI: 0.98, 1.01). Post-UCC, there was a significant increase in the slope of the overall hospital admissions rate of 3.0% (adjusted IRR 1.03; 95% CI: 1.01, 1.04) which is mainly driven by increases in the short-stay admission rate (adjusted IRR 1.08; 95% CI: 1.06, 1.10).

5.7.3.3 Comparison of admission rate Pre/Post UCC summary

The results from this investigation are limited to two UCCs in West London; therefore, generalisability to other UCCs is unknown. The observed effect of the UCC's is highly dependent upon a multitude of inter-related factors that include, the geographical location; the size of service; health profile of local populous; proximity to other healthcare services, e.g. primary care; and closeness or convenience to work for some patients.

An additional factor could be supply induced demand, i.e. the efficiency of the new services in treating walk-in patients may, in fact, cause more patients to choose to re-attend. Moreover, ED services at the HH site have ceased, and on the 10th September 2014, the UCC at HH site became operational 24 hours a day, seven days a week. [345]

Table 5.21 Baseline socio-demographic characteristics of ED attendances and hospital admissions

	Hammersmith			Charing Cross		
	Attends N	Admits N	(Attend ÷ Admit)	Attends N	Admits N	(Attend ÷ Admit)
Total	**278783**	**63127**	**4.4**	**498144**	**111575**	**4.5**
Age group (years)						
0 to 14	12599	67	188.0	14156	9	1572.9
15 to 24	44609	4462	10.0	88150	7940	11.1
25 to 39	95477	12441	7.7	173362	19318	9.0
40 to 59	68244	17020	4.0	113527	27594	4.1
60 to 79	43738	20273	2.2	76420	34138	2.2
80+	14116	8864	1.6	32529	22576	1.4
Sex						
F	158986	34131	4.7	253982	56469	4.5
M	119797	28996	4.1	244162	55106	4.4
Socioeconomic status						
Quintile 1 (most deprived)	105714	22103	4.8	93285	21951	4.2
Quintile 2	92488	20246	4.6	150863	34329	4.4
Quintile 3	50390	12304	4.1	118686	26017	4.6
Quintile 4	19979	5505	3.6	101134	20970	4.8
Quintile 5 (most affluent)	10212	2969	3.4	34176	8308	4.1

Table 5.22 Effect of UCCs on total attendances by hospital site

	Poisson Regression coefficient (95% CI)	
Measure	**Hammersmith**	**Charing Cross**
Time (years)	0.65 (-1.70, 3.00)	0.78 (-1.07, 2.63)
Level change after UCC	11.80 (8.00, 15.60)*	12.77 (9.06, 16.48)*
Slope change after UCC (years)	3.50 (1.03, 5.97)*	0.59 (-1.46, 2.64)
Age group (years) [0-14 reference group]		
15-24	40.20 (37.21, 43.19)*	89.15 (86.12, 92.18)*
25-39	101.50 (98.51, 104.49)*	191.81 (188.79, 194.84)*
40-59	68.69 (65.70, 71.68)*	119.72 (116.70, 122.75)*
60-79	39.16 (36.18, 42.15)*	75.02 (71.99, 78.04)*
80 +	3.18 (0.19, 6.17)*	22.14 (19.11, 25.16)*
Sex [male as reference group]		
Female	-15.96 (-17.67, -14.25)*	-3.94 (-5.69, -2.20)*
Socioeconomic status (Quintile 1 most deprived as the reference group)		
Quintile 2	-13.28 (-15.97, -10.58)*	57.81 (55.05, 60.57)*
Quintile 3	-55.57 (-58.27, -52.88)*	25.50 (22.74, 28.27)*
Quintile 4	-86.44 (-89.14, -83.74)*	7.88 (5.12, 10.64)*
Quintile 5 (most affluent)	-97.13 (-99.84, -94.41)*	-59.35 (-62.11, -56.58)*
Month (January as the reference group)		
February	-3.91 (-8.08, 0.27)	-7.68 (-11.93, -3.42)*
March	1.92 (-2.26, 6.09)	3.97 (-0.29, 8.22)
April	-0.02 (-4.37, 4.34)	0.21 (-4.05, 4.47)
May	2.22 (-1.97, 6.40)	4.70 (0.43, 8.96)*
June	0.42 (-3.76, 4.60)	1.68 (-2.59, 5.95)
July	0.49 (-3.70, 4.67)	4.94 (0.66, 9.22)*
August	-2.11 (-6.28, 2.07)	0.57 (-3.72, 4.86)
September	-1.61 (-5.79, 2.57)	-1.44 (-5.89, 3.00)
October	-0.69 (-4.87, 3.50)	0.63 (-3.63, 4.90)
November	-1.83 (-6.02, 2.35)	-3.94 (-8.21, 0.33)
December	0.23 (-3.96, 4.41)*	-2.46 (-6.74, 1.82)

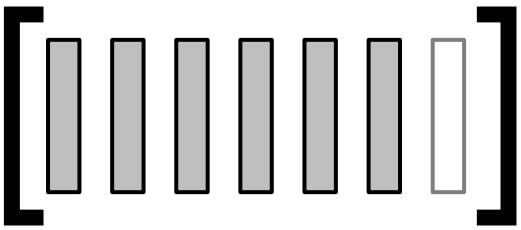
*Significant (P < 0.05) **

Table 5.23 Adjusted multivariate model of effect size of UCCs on emergency hospital admission by hospital site

	Incident Rate ratio, 95% CI		
Emergency hospital admissions	**Annual time trend pre-UCC**	**Level change post-UCC**	**Slope change post-UCC**
Hammersmith			
Short stay (< 2days)	0.92 (0.90, 0.95)*	0.73 (0.70, 0.77)*	1.02 (0.99, 1.06)
Long stay (≥ 2days)	0.98 (0.95, 1.01)	0.90 (0.85, 0.94)*	0.97 (0.94, 1.00)
Total	0.95 (0.93, 0.97)*	0.82 (0.79, 0.85)*	1.00 (0.98, 1.02)
Charing Cross			
Short stay (< 2days)	0.98 (0.96, 1.00)*	0.98 (0.95, 1.02)	1.08 (1.06, 1.10)*
Long stay (≥ 2days)	1.02 (1.00, 1.04)*	0.99 (0.95, 1.02)	0.98 (0.96, 1.00)
Total	1.00 (0.98, 1.01)	0.99 (0.96, 1.01)	1.03 (1.01, 1.04)*

*Significant (P < 0.05) **

Chapter 6 Discussion



6.1 Introduction

This chapter presents a discussion of the results for this study, aligned to meet the specific objectives of this research and reflect the sequential order of processes within the conceptual framework. It begins with a summary of the problem, outlines the research approach and discusses the results for each stage, Selection of informatics techniques, Domain understanding, Semantic interoperability, Syntactic interoperability, Data quality, Data linkage and Linked data analyses, as presented henceforth.

6.2 Summary of the study

6.2.1 Problem

This purpose of this research was to investigate the issue of Health Information Fragmentation (HIF), which is prevalent throughout NHS healthcare services in England. [2] Within the same hospital environment, organisational silos inadvertently develop between individual departments due to differences in specialty, expertise or structural configuration. [4]

Furthermore, each department utilises more than one Clinical Information System (CIS), often designed by different vendors, to facilitate patient management. However, due to a lack of interoperability, these heterogeneous CIS do not 'talk' to one another, i.e. no information exchange; hence, they become isolated. [15, 16]

Correspondingly, patient records transcribed within these disparate CIS are allocated unrelated identifiers, which results in further fragmentation of the patient record, [5] and leads to future complications when trying to re-identify the same patient. [11]

Consequently, the combined effect of departmental silos, disparate multi-vendor CIS and associated technologies, structure/format and heterogeneity of the Electronic Medical Record (EMR) as stored within each respective CIS, inevitably result in 'information' being trapped within data silos. [4, 5]

6.2.2 Research approach

As informed by the literature, the theoretical framework underpinning this research was General Systems Theory (GST), [252] wherein a 'system' consists of interconnected elements that interact to perform a given task, [252, 253, 346] with the additional understanding that the whole system is

greater than the sum of its individual parts. [252, 253] Thus, healthcare environments can be conceptualised as Sociotechnical Systems (STS) encompassing complex issues related to the interconnected nature of 'social' and 'technical' entities, [257, 259, 260] and recursive relations amongst sociotechnical subcomponents. [31]

However, sociotechnical interactions result in unintended consequences that cannot be anticipated at the design or implementation stages. [31] Using an STS lens, if pre-existing ICT infrastructures do not support interaction with newer Health Information Technology (HIT), [31, 32] the lack of interoperability results in the formation of data silos, thus leading to the discovery of the invisible problem of HIF. [30]

Moreover, to permit an in-depth study of the phenomenon of HIF, an Observational Quantitative Single Case Study Research Design was utilised, based on a real-world scenario in which HIF had inadvertently occurred within an Urgent and Emergency Care hospital setting. Whereby, the patient pathway traversed three distinct departments, UCC, ED and IPC, each containing multi-vendor disparate CIS, utilising heterogeneous patient identifiers with no interoperability, ultimately resulting in the fragmentation of the patient EMR and creation of data silos.

The philosophical basis for this research grounded on objectivist epistemology, maintains meaning and meaningful reality exist independently from the operation of any consciousness. [273] Similarly, the issue of HIF exists as an invisible problem [30] regardless of whether anyone is aware of its existence or not.

Adopting a positivist worldview asserts that certain kinds of reliable knowledge can only derive from experience, i.e. empiricism. [251] Thus, using several years of historical routinely collected administrative data from a single case, this research investigated HIF, without interference from the author, to identify its causes and consequences, and permit subsequent application of select informatics techniques that enabled hypothesis testing. [269]

Furthermore, this research took a deductive approach to theory development, focusing on record linkage theory, moving from general premise to a particular situation and conclusion [250] using a quantitative research design. [280] Thus, numerical data was utilised to test empirically the hypothesis that employing deterministic linkage or probabilistic linkage strategies would yield different results.

A case study is “an empirical method that investigates a contemporary phenomenon (the ‘case’) in-depth and within its real-world context”. [290] Moreover, as Emergency Medicine is a sociotechnical process [262] and data silos are considered sociotechnical problems [28], the adoption of a single case study strategy based within an Emergency Care Hospital setting faced with the issue of data silos, presented an ideal opportunity to study in-depth the phenomenon of HIF.

Thus, following a comprehensive review of the literature, a conceptual framework employing appropriate methodologies was created to address the ‘wholeness’ of the HIF phenomenon, providing an end-to-end approach, beginning with multiple heterogeneous data silos and ending with the creation of a Linked dataset, immediately useable for undertaking analyses not previously possible with unlinked data.

6.3 Discussion of the results

6.3.1 Selection of informatics techniques

Research Question 1.	
What informatics techniques address the issue of health information fragmentation?	
Research Objective:	
2	To identify informatics techniques that enable the linkage and integration of heterogeneous data sources.

To answer the first question and associated objective of this research, a comprehensive literature review revealed that although prior studies have noted the importance of addressing HIF, ‘data integration’ is not just one problem, nor is there one unifying theory of integration, which corroborates with the findings of Haas (2007), [27].

Each ‘integration’ scenario presents a specific challenge that is explicitly dependent upon the environment in which HIF has occurred. However, the most important challenge lies in the correct identification, selection and combination of tools, technologies and processes used to perform integration. [27]

Although the issue of HIF is complex and multifaceted, the majority of studies identified in this research have solely focused on particular aspects of HIF. Nevertheless, it is possible to overcome HIF by focusing on and tackling each facet individually, via the identification and utilisation of an appropriate theory, methodology or informatics technique aimed at addressing that specific facet.

Correspondingly, this research viewed the issue of HIF as one large issue composed of interlinked smaller issues and took a holistic approach considering the entire integration process from end-to-end, i.e. beginning with multiple data silos and ending with a linked dataset.

Thus, a holistic view combined with a systematic and logical approach informed by the literature enabled the alignment of informatics techniques, evidenced by their use in prior research, to facilitate the creation of a conceptual framework as used within this thesis, to achieve data linkage and integration. Furthermore, an important factor was the sequence in which the identified informatics techniques were applied, making them 'fit together' as the results from each step contributed to the next step.

To achieve end-to-end integration, this research did not rely on any particular software application; rather, it focused on using techniques identified within the literature. One of the main advantages of using existing theories, methodologies and informatics techniques is that they have been developed and adapted by researchers within the field over many years.

Additionally, the use of such methodologies has been evidenced and documented within many different study settings, using different applications, demonstrating wide applicability, and researchers will know from the onset exactly what outcome a particular technique will achieve. Ultimately, factors such as study purpose, data availability and research outcomes will determine which techniques are to be utilised for any piece of research.

Research Question 2.	
How can identified informatics techniques be adopted to support the linkage and integration of disparate data silos within a hospital setting?	
Research Objectives:	
1	To explore the formation of data silos within healthcare settings.
3	To apply techniques that enable interoperability between the clinical information systems in the emergency care setting.
4	To select data warehouse architecture appropriate for housing healthcare data extracts
5	To determine if data are fit for linkage purposes by assessing the quality of data.
6	To differentiate the effect of deterministic and non-deterministic data linkage strategies with respect to the numbers of correctly matched records.
7	To calculate the transition time for patient transfer between urgent and emergency care services
8	To distinguish the effect of utilising Linked and Unlinked datasets upon the ED 4 hour standard.
9	To compare the rate of emergency patient admissions pre/post UCC implementation.

Moreover, the second research question focused on the application of the identified informatics techniques to support linkage and integration of disparate data silos. As HIF is multifaceted, each objective was devised to target a particular facet; hence, these objectives and associated techniques are discussed in the next section.

6.3.2 Domain understanding

Addresses objective 1: To explore the formation of data silos within healthcare settings.

This purpose of this objective was to enable the author to gain an in-depth understanding of the patient pathway throughout the UCC, ED and IPC departments and identify the capture and flow of information, and any interrelationships between the three CIS. This helped determine the extent to which information was being fragmented and led to the identification of factors contributing to the formation of data silos.

As mentioned in the literature, approaches such as Business Process Modelling (BPM) permit the creation of visualisations that represent organisational workflow [308, 309] using Unified Modelling Language (UML) to model domains. [310] The application of these approaches permitted the creation of a UML Activity diagram and a Use case diagram, while further CIS training and collaboration with vendors and domain experts led to the identification of heterogeneity amongst the three CISs.

Within the study setting, data silos were formed due to the introduction of new HIT that was not interoperable with existing technologies, which aligns with the sociotechnical problem of unintended consequences and emergence of the invisible problem of HIF within an organisation. [30-32]

Additional factors contributing to silo formation include services working in isolation, the heterogeneity of CISs hosting data [23] due to multiple vendors, disparities in EMR structure, and the co-location of modern CISs with legacy CIS [12] preventing interoperability due to technological differences.

However, a very limited amount of interoperability was identified, as only the IPC database was able to transmit an electronic message (one-way) to the ED database, containing only demographic details and a temporary hospital identification number.

Furthermore, patients seen within the UCC, ED and IPC departments were allocated three heterogeneous IDs, respectively, which added a further layer of complexity when trying to correctly identify patients across departments.

Moreover, the ED database remained isolated as first; it did not have direct access to the NHS Personal Demographic Service and relied solely on the IPC database message to populate the NHS number field. Second, the imported temporary IPC hospital ID number automatically became redundant when administrative staff retrospectively allocated permanent patient IDs, but this updated ID was not fed into the ED database.

A surprising finding was 'Paper persistence' as highlighted in the literature, [347] when services faced with a lack of interoperability between CISs, they rely upon using paper to 'bridge the gap'. This characteristic is reflected in the study setting, in which patients are handed paper forms for either undertaking diagnostic tests or referral from UCC to ED. Due to a lack of intra-hospital data sharing, the burden of information gathering and transfer often falls upon clinicians and patients themselves. [348]

Correspondingly, another contributing factor to data silo formation was the use of paper 'Casualty Cards' in ED, whereby handwritten notes were transcribed onto a paper template to be later scanned and attached to the patient record, rendering such data completely inaccessible. [5]

These findings provide some explanation and offer an insight into the underlying reasons as to how data silos are inadvertently formed and emphasise the invisibility of the problem. [30] The reasons are a complex mixture of social factors, i.e. how departments function operationally, actor interactions, and technological factors: multi-vendor, modern or legacy systems, and a lack of utilisation of technology, e.g. use of paper records.

Thus, the extent of sociotechnical issues becomes particularly apparent during change processes, whereby the implementation of new HIT can influence/impact existing relationships. [260] The findings from this case study thus far, align with the literature and present evidence that unequivocally places the issue of HIF and data silos within the sociotechnical domain.

Nonetheless, the most important factors when attempting to undertake domain understanding is identifying, interpreting and combining the knowledge held by domain experts, clinical, managerial,

and administrative staff, and system vendors, along with findings from within the research literature to understand concepts. Although BPM and UML do not offer a solution to HIF, they are tools that enable researchers to illustrate their perspective on HIF within a particular setting.

In summary, although HIF and data silos result from unintended sociotechnical consequences, services implementing new HIT, or simply reviewing existing technologies, should consider the benefit of undertaking 'domain understanding' processes to document/highlight contributory factors leading to HIF. The discoveries from undertaking such processes will serve to inform researchers which particular integration approach to adopt, e.g. as the CISs within this study were not interoperable, the literature had specifically identified CDW techniques to overcome this issue.

6.3.3 Semantic interoperability

Addresses objective 3: To apply techniques that enable interoperability between the clinical information systems in the emergency care setting.

Having identified multiple disparate CISs, the author was faced with the problem of not knowing whether the data contained within the three CISs shared the same precise meaning, i.e. were they semantically interoperable? The level of heterogeneity amongst CISs will strongly influence the properties, formats, types and values of data contained within each CIS, and data provenance may be unclear. [177-179] Hence, this objective sought to utilise techniques that enabled the correct identification and standardisation of specific data elements required for data linkage and integration.

As mentioned within the literature, ontologies provide a conceptualisation of the structured representation of a given domain or meaning [146] and provide a method for describing concepts and relationships within that domain. [147] As conceptualisations are not bound to any one specific system, they can be used to identify commonalities amongst disparate data sources.

Within this study, as each of the three CISs utilised different naming conventions, the first step to overcoming heterogeneity was by creating a 'Patient Ontology' model based upon the concept of a patient who is a person with attributes including name, gender, birth date, address, and patient ID. [149, 311] Using the patient ontology to cross-reference the respective data dictionaries for each of the three CISs enabled the correct identification of data elements required for linkage and integration.

Furthermore, the standardisation of data elements as per the NHS Data Dictionary [311] or Fast Healthcare Interoperability Resources [160] serve to ensure the precise meaning of information is understandable by other systems. [131]

Thus, the second step involved the collation of metadata and the creation of bespoke reports within each CIS, designed to enable standardisation. Data element naming conventions were standardised by allocating a common pseudonym, and data conversion scripts ensured the selected data elements possessed exactly the same data type, format and properties, to ensure consistency across the three CISs to enable subsequent linkage and integration. For example, dates formatted as DD/MM/YYYY or YYYY/MM/DD would not support linkage.

As expected, multi-vendor CISs utilise dissimilar naming conventions, some of which are not meaningful. Surprisingly, the design and level of detail across the three data dictionaries varied considerably, with one dictionary not accurately reflecting the live implemented CIS. The latter was due to vendors creating multiple versions based on one central database model, adapted for use in different settings.

However, these adaptations were not fully documented, which proves challenging for researchers when attempting to identify specific data elements and necessitates seeking advice from database managers. For researchers, to avoid misinterpretation one potential approach to mitigating this issue would be to undertake system-specific training, as this permits a greater understanding of the nuances of each CIS.

Furthermore, as CISs are designed for use within healthcare settings, vendors should consider using naming conventions and attributes that conform to NHS Data Standards, [311] otherwise the tasks of data interpretation and applying conversions fall upon researchers using the data.

Overall, these findings highlight the importance of considering semantic interoperability when undertaking data linkage and integration tasks and demonstrate the applicability of ontologies, which permit users to specify the level of detail required and standardisation approaches, e.g. based on NHS recommendations. Such approaches are not system-specific and provide a means for researchers to overcome heterogeneity across disparate datasets.

6.3.4 Syntactic interoperability

Addresses objective 4: To select data warehouse architecture appropriate for housing healthcare data extracts.

Although the previous objective enabled the identification and creation of bespoke reports to standardise the data elements required for this study, nevertheless, data remained trapped within each respective CIS. Therefore, this objective sought to utilise informatics techniques that enabled data from each of the three CIS to be brought together for linkage and integration.

However, one of the key issues with multi-vendor CISs is heterogeneity due to system design, platforms and interfaces, [23, 118] that act as barriers and prevent any form of data transfer between systems. Although vendors have specific expertise in the domain for which they design their CIS, they may not collaborate to facilitate interoperability, for fear of divulging details, rival competition, intellectual property, time investment, costs and adapting their own systems to accommodate connections with other systems, consequently leaving data within silos. [12]

Fundamentally, syntactic interoperability focuses on the process of moving data between two systems, i.e. the ability of computer systems to exchange data, without necessarily understanding what is being exchanged. [7, 130]

Thus, to overcome this impasse, the literature identified approaches including Knowledge Discovery in Databases (KDD) [167, 168], and sequential data processing stages envisioned as a 'Data Pipeline' to provide guidance on the overall process. [169] Additionally, CDW (or OLAP) technologies are specifically designed to facilitate data integration from multiple heterogeneous systems, via the importing of numerous data extracts into one CDW for subsequent linkage and integration. [171]

Additionally, CDW technologies exist independently of any CIS and are not restricted or affected by system nuances nor vendors. Researchers have the freedom to select appropriate architectures, e.g. Top-Down (Inmon) [173] or Bottom-up (Kimball) [174] and design a CDW as considered appropriate for their study.

Furthermore, as the three CIS were not syntactically interoperable, this study followed the principles of KDD and Data Pipelines, and used Top-Down CDW architecture to obtain data extracts from the UCC, ED and IPC databases, and store these within one location, i.e. the CDW.

Moreover, an important consideration when using CDW approaches is the underlying ICT infrastructure. Studies utilising healthcare data must ensure that data remain within a secure environment, with restricted access, adhering to the host organisations Information Governance policies and procedures.

In addition, the CDW Database Management Software utilised within a study will be dependent upon what the host organisation typically use, e.g. Microsoft SQL Server for this study. Commonly within healthcare settings, ICT departments facilitate routine maintenance tasks, e.g. implementing security and undertaking regular server backups.

As identified in this study, the CIS's server storage locations can vary; two of the CISs were located locally within the hospital setting, which permitted seamless data extraction and import into the CDW. However, one vendor stored their CIS data externally to the hospital, only providing access via a web interface, so syntactic interoperability was not possible. Furthermore, due to the time frame for this study, large data file sizes and limitations on bandwidth, data extracts were obtained as individual monthly files via secure file transfer protocols, which proved time-consuming.

In summary, the results demonstrate the advantages of using CDW technologies guided by KDD and Data Pipeline approaches to facilitate syntactic interoperability, and address another important facet of HIF. Although data has been extracted from multiple CIS and stored within one CDW, the question of how to link the datasets remains unanswered; however, this was addressed by subsequent objectives.

6.3.5 Data quality

Addresses objective 5: To determine if data are fit for linkage purposes by assessing the quality of data.

When considering Data Quality (DQ), a crucial question that should be asked is "are these data fit for the purpose intended?". [184] Thus, data can be considered to 'have quality' if they satisfy the requirements of their intended use, and to achieve this data must be trusted and understood. [188]

This study utilised routinely collected hospital patient data, obtained from the CISs within an emergency care setting. However, due to the hectic nature of ED, and heterogeneity amongst CISs,

data may contain inconsistencies, be inaccurate, incomplete, uninterpretable, hence, are corrupted and unlikely to be of benefit, rendering such data unsuitable for research. [98, 112, 177-179]

Ultimately, the quality of the underlying data will influence the strength of any type of research or evaluation. [93] Therefore, prior to undertaking linkage and integration processes, this objective sought to assess the quality of each data element within the UCC, ED and IPC datasets.

To achieve this, the literature identified several DQ dimensions, i.e. the degree to which data items were complete, valid, unique, accurate, consistent, timely, sufficiently detailed, and appropriately represented whilst retaining sufficient contextual information to support the decision-making process. [26] Nonetheless, the study purpose, specific data elements to be assessed, and researcher's preference determine the specific dimensions to be used for DQ assessment.

For this study, the level of DQ for each data element within all three datasets was assessed and compared. Not surprisingly, the episode and patient identification numbers possessed 100 per cent completeness, as these were automatically created by each CIS at the start of a new patient presentation. However, as they were all heterogeneous, they were unsuitable for linkage purposes.

Additionally, the completeness for the NHS number within the ED dataset was approximately 45.8%, which was relatively poor compared to the UCC (78.5%) and IPC (94.8%). This was, however, due to the ED database not possessing the functionality to connect to the NHS Personal Demographic Service at patient registration; hence, the NHS number was also found to be unsuitable for data linkage.

Nonetheless, the literature had identified data linkage techniques using quasi-identifiers such as an individual's demographic information: forename, surname, date of birth, gender and postcode. [202] Following DQ assessment and error correction, the quasi-identifiers from across all three datasets had a minimum of 99.9% completeness, except postcode, which had a minimum of 96.6% completeness, thus, confirming their suitability for data linkage.

As the patient name data elements were free text, they contained the most inconsistencies and errors, requiring identification and correction, e.g. inclusion of spaces, special characters, numbers, 'Unknown' and even Date of birth entries. Most surprisingly, one of the data extracts contained 'Test' or 'Dummy' records that should be have been removed by the vendor prior to live

implementation of the CIS, and for Date of Birth default dates, i.e. '01/01/1900' or '01/01/1901' were addressed.

When working with healthcare data, ensuring good DQ is crucial, e.g. a misplaced decimal point will make a significant difference to medication doses. Although some CIS include DQ reports, often these focus on a limited set of DQ dimensions, e.g. completeness or uniqueness. Ideally, vendors should consider collaborating with consumers to identify key metrics and develop DQ reporting tools within the CIS that enable data to be checked periodically to permit the identification and correction of errors.

The results from this study demonstrate the importance of undertaking DQ assessments and data cleaning tasks to enable researchers to understand, have confidence and trust their data. [188] DQ assessment is a crucial component of HIF as it prepares the data for subsequent linkage and integration.

Thus, for research studies, the selection of appropriate DQ dimensions depends as much on the intended use as it does on the data itself. [188] Researchers must also consider how they will deal with inconsistent data and how they define and measure DQ. [27]

Moreover, the five paradigms Data, Information, Knowledge, Understanding and Wisdom, [100] begin with data. Thus, if data are erroneous, it follows that any conclusions drawn or wisdom acquired will undoubtedly be unreliable.

6.3.6 Data linkage

Addresses objective 6: To differentiate the effect of deterministic and non-deterministic data linkage strategies with respect to the numbers of correctly matched records.

The objectives and techniques thus far have enabled the identification and standardisation of specific data elements from multiple CIS, the extraction of datasets into a CDW and an assessment of DQ, all of which serve to prepare data for linkage. However, the discovery that disparate datasets lack a common unique identification number, poses the question, how do we link and integrate the three datasets?

Record Linkage (RL) brings together information from two records believed to relate to the same entity, e.g. an individual. [197] The literature identified Deterministic (DRL) and Probabilistic (PRL) linkage techniques, along with a range of complementary techniques to enhance the linkage process.

Thus, using exactly the same datasets within the CDW, the purpose of this objective was to apply and differentiate between DRL and PRL strategies, to deduplicate patient records, integrate the three datasets, and ultimately overcome the issue of HIF by creating a linked dataset, as per research question two of this thesis.

First, DRL was achieved via the utilisation of the HES algorithm, which is based upon a 'Step-based DRL' strategy [206] using the NHS Number and patient quasi-identifiers date of birth, gender and postcode, and was applied to the three datasets.

Second, informed by the literature, the author developed a Hybrid algorithm [201, 220, 244] which utilised both DRL strategies to deduplicate via ID numbers and PRL strategies based on the Fellegi-Sunter [209] approach, using quasi-identifiers: surname, forename, gender, date of birth and postcode.

Additionally, the PRL stage of the Hybrid strategy incorporated complementary techniques to enhance the linkage process, listed henceforth. Blocking to reduce the number of comparison pairs and increase efficiency [198] via alphanumeric Soundex codes (name phonetic algorithm). [199] Using quasi-identifiers forename, surname, gender and date of birth to create a Combination ID. [202] Splitting the date of birth, into its component parts Day, Month and Year [198, 213] to allow for transcribing errors. The use of Jaro-Winkler string comparator [199, 202, 229, 230] to account for insertions, deletions, and transpositions between two strings. The calculation of Relative frequencies [199, 213] for both rare and common data items as found in surname and forename, to permit value-specific frequency-based modification.

Overall, objective six sought to differentiate empirically whether the application of a particular strategy DRL (HES strategy) or PRL (Hybrid strategy) influenced the number of correctly identified records belonging to the same individual across multiple datasets. The findings show the HES algorithm linked approximately 711 thousand records, whereas the Hybrid algorithm was able to link 1.07 million records, the difference between the two strategies was statistically significant ($P<0.001$).

An explanation for this difference is that the HES algorithm is dependent upon the NHS number to undertake data linkage, and poor recording (45.8% completeness) of the NHS number within the ED dataset limits the overall number of records that can be linked. However, the Hybrid algorithm was not dependent upon the NHS number and relied upon the Quasi-ID, created from patient demographic details which recorded over 99.9% completeness, resulting in a much higher number of linked records.

Furthermore, the HES algorithm is used by NHS England to link patient-level hospital data, including ED attendances in England. [206, 207] This distinctive feature of the algorithm made it ideal for application within this research, as this study utilised ED datasets. Therefore, comparing the results from the HES algorithm with the Hybrid algorithm provides a means of validation, to confirm that the results from the Hybrid algorithm are indeed credible.

Although the results from this study cannot be directly compared with results from other studies, the findings that PRL strategies outperformed DRL strategies were consistent with the findings by authors: Kabudula et al., (2014), [220], Méray et al., (2007) [223], Sayers et al., (2015), [199], and Dean et al., (2001). [216]

Additionally, the Hybrid linkage strategy developed and applied by this study differs from the Hybrid strategy utilised by Kabudula et al., (2014), [220], who first used a DRL strategy to link two datasets, then used a PRL strategy for the remaining unlinked records. An issue with this approach is that the author did not attempt to link the PRL group to the DRL group, thereby missing any potential matches resulting from the presence of typographical/transcribing errors, e.g. comparing Grey with Gray.

Nonetheless, one of the limitations of the PRL approach is that the Match and Non-match probabilities are not known beforehand, nor can they be obtained from within the literature. Ideally, they need to be calculated from the source data used within a research study. Although some of the studies identified in the Literature Review, e.g. Kabudula et al., (2014), [220] use the Expectation Maximisation algorithm, it does not always yield reasonable parameters and researchers must exercise caution when utilising it with real-world data. [197] To mitigate this issue, this study calculated the Match and Non-match probabilities from the datasets within the CDW to ensure accuracy.

Moreover, a surprising finding was the effect of transposing or reshaping postcode data from long to wide. For example, one patient record contains six data elements, of which five are exactly the same, i.e. ID, Surname, Forename, Date of Birth and Gender, however, the same patient has four associated postcodes. Therefore, within a database, the values for the first five data elements will repeat four times, with only the postcode value being different each time.

Consequently, this will result in a superficial increase in the number of comparison records when undertaking PRL. Thus, within the main dataset, there were N=635,069 unique records pre transposition, which fell to N=615,369 post transposition, a reduction of 19,700. For example, applying a Cartesian product to only the reduction value returns 19,700 x 19,700 = 388,090,000 or approximately 388 million superfluous record comparisons that were avoided due to transposing the postcode data element.

Importantly, the selection and application of the Hybrid algorithm resulted in the creation of one Master Patient Table containing records for every individual patient across all three UCC, ED and IPC datasets, uniquely identifiable via the Quasi-ID. This table was designed/structured to reflect the Electronic Health Record, providing a longitudinal record of all patient attendances and admissions chronologically. Each patient was identified uniquely via the Quasi-ID, which in turn linked to each Patient Episode and subsequent Clinical Encounter [344] across the UCC, ED and IPC departments covering both hospital sites.

Furthermore, The Master Patient Table facilitated the connection of the disparate datasets to create one Linked dataset, which did not contain any duplicate records. Thus, addressing the issue of double-counting and overestimation of patient numbers, i.e. those patients referred from the UCC to ED and vice versa that were being counted twice due to having EPRs within both CISs.

In summary, the principal finding was that there was no ‘one’ superior linkage strategy, rather, the selection and adoption of particular strategies are determined by first, the purpose of linkage, i.e. to simply merge/join tables or undertake deduplication (entity resolution) processes. Second, the availability and presence of common data elements within each of the datasets, and third the preferences of the researcher, based upon their level of knowledge and experience. Overall, the informatics techniques utilised thus far have enabled the complex web of HIF to be disentangled, resulting in the successful creation of a Linked dataset that permits novel analyses to be performed, as discussed in the next section.

6.3.7 Linked dataset analysis

The aim and objectives for this research have culminated in successfully overcoming the issues of HIF and data silos, resulting in the creation of a linked dataset. However, the creation of a new dataset does not necessarily constitute new knowledge. Data must first be analysed and interpreted before it can be said to have yielded fresh insights and understanding. [33]

An important consideration is the purpose of linkage, what do researchers intend to do with linked data? Therefore, this section serves to provide evidence of the benefit and practical applicability of the linked dataset by demonstrating its use via the presentation of three examples of analyses that were not possible to undertake using fragmented datasets.

6.3.7.1 Transition times

Addresses objective 7: To calculate the transition time for patient transfer between urgent and emergency care services

Patients initially seen in the UCC and subsequently referred to ED were given a referral letter and had to queue at reception to register for ED services, this additional waiting time was calculated as the Transition time, which is the time in between departmental patient transfers that is not captured within any one database system. However, it contributed to the total amount of time a patient spent in the department and was affected by busy periods. The linked dataset permitted the calculation of the transition time, and this was found to vary by site, e.g. the mean time at HH was approximately 13 minutes and CXH approximately 7 minutes.

Importantly, the transition time reflects a lack of interoperability between the UCC and ED databases and permits the measurement of its effect, impact and cost in units of patient time (minutes). If the two CISs could communicate and send an electronic message of patient details from the UCC to ED database, patients would not have to spend this extra time waiting to re-register. Additionally, as this time is not measurable using unlinked data, it cannot be included within reporting metrics, i.e. the total time a patient spends within the ED setting.

This finding demonstrates how the issue of HIF reveals itself, using paper to the bridge the gap of fragmentation costs time and unnecessarily prolongs the pain or suffering experienced by patients waiting to receive treatment.

6.3.7.2 ED four-hour standard

Addresses objective 8: To distinguish the effect of utilising Linked and Unlinked datasets upon the ED 4 hour standard.

The four-hour standard is a Clinical Quality Indicator [68] focusing on the total time a patient spends within ED (minutes) from Arrival time to Depart time (including admission to IPC), and a key indicator is the 95th percentile. Any patient record in which the total time in ED is greater than four hours is considered a 'Breach' and this record requires retrospective clinical investigation.

Focusing on patients referred from ED to IPC, the linked dataset identified and resolved chronological discrepancies between the date/time as recorded within each respective CIS, whereby the ED episode End date/time was greater than the IPC episode Start date/time.

Furthermore, this study demonstrated the benefit of the Linked dataset versus an unlinked ED dataset, by calculating/comparing the metrics ED time in department, and four-hour standard, as proxy indicators. The results identified a statically significant difference ($P<0.0001$) between the Mean times, for CXH site, Unlinked dataset Mean 229.14 minutes (95% CI: 228.26, 230.03) and Linked dataset Mean 153.42 minutes (95% CI: 153.03, 153.82), the findings were similar at the HH site. These findings suggest that using the Linked dataset provides a genuine reflection of the four-hour metrics, whilst reducing the number of breaches that require clinical review, saving valuable clinical time.

Moreover, this example clearly demonstrates the benefits associated with using the Linked dataset compared to using an unlinked dataset, to obtain accurate results. In this context, a lack of interoperability between two CISs can be measured in units of Clinician time unnecessarily spent reviewing four-hour breaches that were not clinical in reason, but nevertheless flagged due to inaccurate time recordings within CISs.

6.3.7.3 Comparison of admission rate pre/post UCC

Addresses objective 9: To compare the rate of emergency patient admissions pre/post UCC implementation

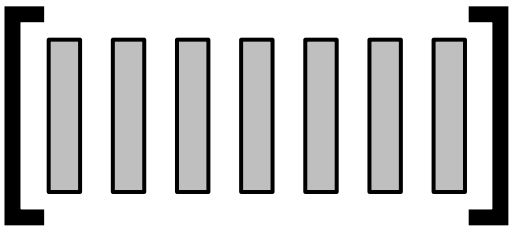
The study time frame spanned seven calendar years from 1 January 2007 to 31 December 2013, during which the GP-led UCCs were implemented in mid-2009 at both HH and CXH sites, respectively. "How do we know if change is an improvement?" [337] To answer this question, the Linked dataset was used to look at the rate of emergency patient admissions pre/post UCC implementation by undertaking Interrupted Time Series (ITS) analysis.

The Linked dataset encompassed all patient attendances and admissions from the UCC to ED to IPC, and previous linkage strategies ensured that there were no duplicate records present, and individual patients could be identified across the entire patient pathway throughout all three departments.

This section permitted comprehensive analysis of the effect of the UCC at both sites, for HH site there was found to be a significant reduction ($P<0.05$) of admission rates post-UCC, whereas this was not the case at CXH. However, both sites showed an increase in the number of patient attendance which may suggest supply induced demand. Importantly, ED services at the HH site ceased operation on 10th September 2014, and the HH UCC became operational 24 hours a day, seven days a week.

Moreover, as research informs decisions, researchers must have confidence and trust in their data. [188] The results from the ITS analysis would not have been possible using unlinked datasets, which again highlights the distinct advantages and benefits of using DQ-checked linked datasets to undertake in-depth analyses that can aid service evaluation and inform decision-making processes.

Chapter 7 Conclusion



7.1 Conclusion

This thesis presented an original perspective on the phenomenon of HIF and argued that although a unifying theory of data integration and exchange has yet to be established [27], the issue of HIF and data silos can, nevertheless, be addressed by using selective informatics techniques identified within the research literature.

The study set out to answer two questions, first, what informatics techniques address the issue of HIF, and second, how can these identified informatics techniques be adopted to support the linkage and integration of disparate data silos within a hospital setting.

Therefore, to examine the impact of HIF, this study utilised a single case study research design based on a real-world scenario faced by the UCC, ED and IPC services within a large metropolitan London hospital, in which routine administrative data were held in silos within three disparate CISs.

Nevertheless, each integration scenario presents a specific challenge that is explicitly dependent upon the environment in which HIF has occurred. However, data integration is not just one problem, and from a theoretical perspective, there remains a lack of understanding in terms of what fundamental processes are capable of integrating information. [27]

Moreover, the most important challenge lies in the correct identification, selection and combination of informatics tools, technologies and processes used to perform integration. [27] Therefore, to answer the two research questions of this study and demonstrate that, although there is no unifying theory of data integration [27], several techniques may be 'fitted together' in specific combination to achieve data integration. This principle has been envisioned as a conceptual conclusion, based on the findings of this study, and presented in Figure 7.1 (page 275).

7.1.1 Problem classification (Figure 7.1)

General Systems Theory (GST) maintains a 'system' consists of interconnected elements that interact to perform a given task. [252, 253, 346] Encompassed within GST lies the Sociotechnical Systems (STS) perspective, whereby, healthcare settings can be conceptualised as heterogeneous systems [257] in which all actors (human and non-human) are interconnected and interdependent on each other to perpetuate action, [32] for example, Emergency Medicine is a sociotechnical process. [262]

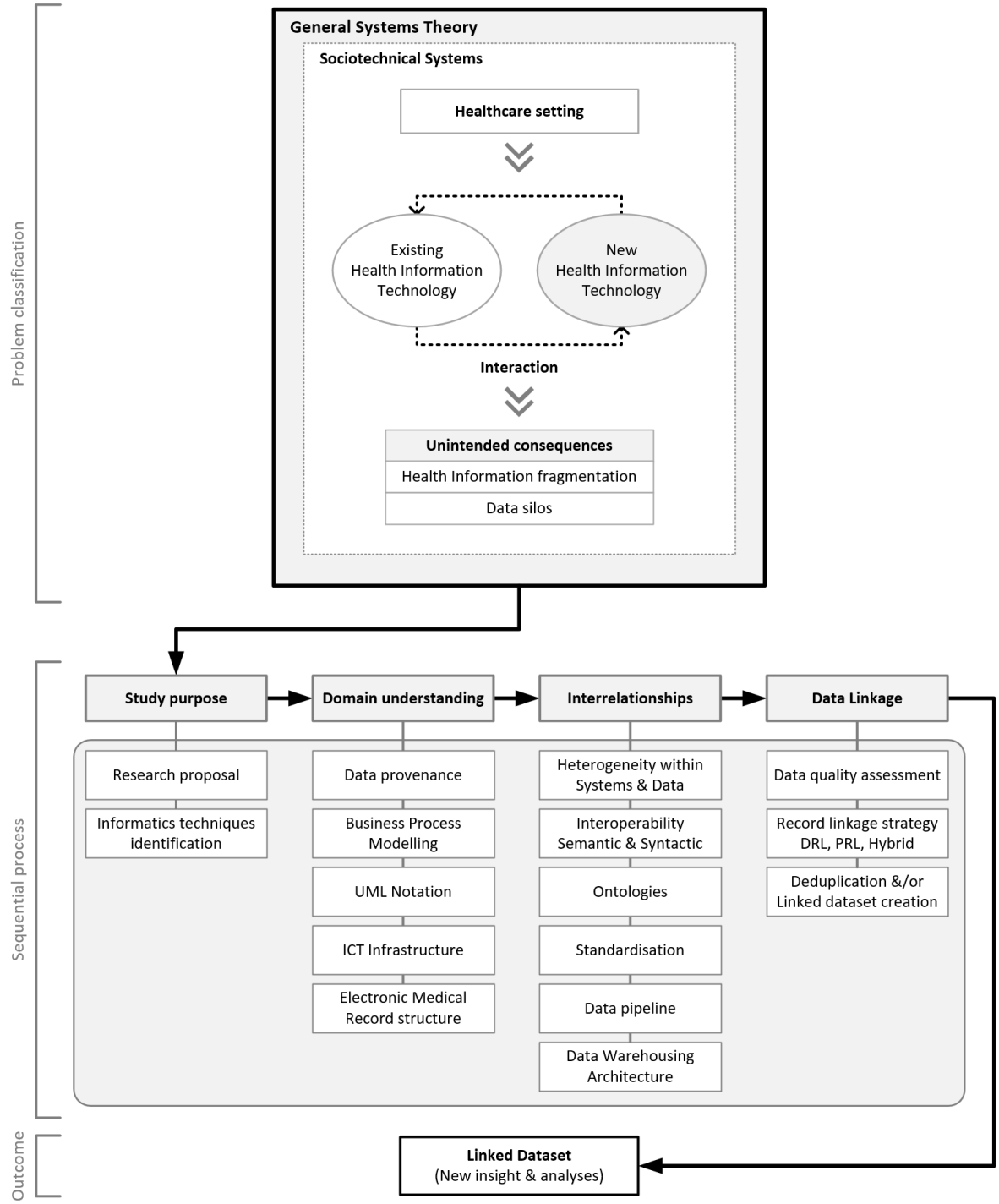


Figure 7.1 Conceptual conclusion based on research findings

Furthermore, it is the recursive interrelationship and technical interaction between sociotechnical subcomponents, i.e. for data exchange, that affects the degree of interoperability between pre-existing and introduction of new HIT. [31, 32, 257] Sociotechnical interactions result in unintended consequences that cannot be anticipated at the design or implementation stage. [31]

Therefore, a lack of interoperability, in which subcomponents do not 'talk' to one another leads to system isolation [15] and the discovery of the invisible sociotechnical problem of HIF resulting from data silos within an organisation. [28, 30-32]

7.1.2 Sequential process (Figure 7.1)

Achieving interoperability is just one of the many STS challenges faced by healthcare organisations. [257] This research took the view that although the issue of HIF is complex, it is, however, multifaceted and can be subdivided into a series of more specific interlinked issues. Thereby taking a holistic approach and considering the entire integration process from end-to-end, i.e. beginning with multiple data silos and applying specific informatics techniques, will result in the creation of a linked dataset.

Although the literature identified specific informatics techniques, each aimed at addressing a particular facet of HIF, the most important factor was determining the sequence in which the techniques were applied, making them 'fit together', as the results from each step contribute and inform the next step.

First, the study purpose will justify the requirement for the use of linked data in meeting study objectives and outcomes. Additionally, it will incorporate an assessment of potential informatics techniques for data linkage and integration purposes.

Second, domain understanding and associated techniques facilitate researchers in identifying HIT and discovering the underlying causes and extent of HIF within the study setting. In addition, this knowledge will enable researchers to make informed decisions when selecting the most appropriate informatics techniques for their study.

Third, interrelationships, by considering the extent of heterogeneity amongst multiple HIT and data contained within, researchers can utilise informatics approaches that permit the correct identification, standardisation and storage of data elements, as required for their study.

Four, data linkage processes begin with an assessment of data quality, via specific quality indicators and subsequent data cleaning. The selection of an appropriate data linkage strategy is entirely dependent upon the presence/absence of common unique identifiers throughout multiple study

data. Additionally, the purpose of data linkage, i.e. to perform deduplication or creation of a linked dataset will have been pre-determined by the study purpose.

7.1.3 Outcome (Figure 7.1)

In summary, the careful consideration of each process sequentially will enable researchers to overcome HIF. Moreover, the conceptual conclusion for this study aligns with GST, which asserts the whole system is greater than the sum of its individual parts. [252, 253]

Correspondingly, this study has demonstrated the amalgamated effect of utilising informatics techniques sequentially is far greater than using these techniques individually, as the combined effect successfully overcomes HIF and data silos, resulting in the creation of a linked dataset that facilitates new insights and analyses.

Moreover, data must first be analysed and interpreted before it can be said to have yielded fresh insights and understanding. [33] To address this statement, disparate routine administrative data, were given entirely new qualities via their integration, and an additional dimension via the creation of a linked dataset, which following analyses, revealed new insights and created new knowledge, not previously possible using unlinked data.

7.2 Concluding comments

This thesis has presented an example of how the problem of HIF and data silos can be overcome using data integration and linkage techniques within a real-world hospital setting. However, data integration remains a challenging task, explicitly dependent upon the environment in which it has occurred. It is not just one problem, and there remains a lack of understanding in terms of correctly identifying the fundamental processes, associated theories, and technologies capable of integrating information. [27]

Although some researchers envisage moving beyond individual integration techniques to building a single integration engine that automatically selects the right techniques for a given situation, [27] to accomplish this in reality, the proposed integration engine would require awareness of every conceivable integration scenario.

Until such a time, this study has demonstrated the effectiveness of fitting together existing methodologies in a sequential manner to successfully overcome HIF. Although this study covers one integration scenario, there remain many other techniques and methodologies applicable to different integration scenarios, requiring research, exploration, investigation and application.

Nonetheless, the problem of HIF will exist as long as health services remain fragmented. Thus, researchers are encouraged to continue investigating this issue, develop strategies based on recognised methodologies and theories, to develop and test novel approaches and strategies to utilising and applying informatics techniques to achieve integration and linkage, inform the wider scientific community and support the progression of knowledge.

Where is the life we have lost in living?

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?

T.S. Eliot, The Rock (1934)

7.3 Implications

Policy-makers, service managers and researchers need better methods to measure and understand the existing conditions that underlie a lack of interoperability within hospital settings, [348] as HIF is dependent upon the environment in which it occurred. This study demonstrates the benefits of undertaking 'domain understanding' processes to document and highlight contributory factors leading to discovering the extent of HIF, to make informed decisions on how best to tackle the issue.

Furthermore, this study has demonstrated that record linkage without any common patient identification number is possible using patient quasi-identifiers and novel PRL techniques, to overcome the issue of patient identification amongst heterogeneous data. Such techniques facilitate implementation in a variety of healthcare settings and may be potentially useful tools, e.g. for hospital analysts and researchers.

Within EDs, CISs are designed such that the EPR is based on a timestamp, so any historical ED attendance, treatment and associated details will not be available to clinicians. For patients that repeat attend ED, the ability for clinicians to be able to view details of previous attendances would lead to more informed decisions about courses of action and treatment plans. Having access to a resource such as a CDW linked dataset would serve to mitigate this situation, enabling ED clinicians to view past attendances and outcomes to aid patient treatment.

Furthermore, CIS vendors are advised to consider conforming to health interoperability standards when creating systems, e.g. referring to NHS Data dictionaries, HL7 FHIR, to facilitate a common language, terminology and specification that enable consistency between different systems and the data contained within, i.e. semantic interoperability. Furthermore, vendors should consider incorporating dedicated 'space' to store patient IDs from other CISs to assist analysts with correct patient identification across disparate CIS.

7.4 Contribution to knowledge

7.4.1 Informatics techniques

This is the first study to consider the issue of HIF as multifaceted, i.e. one large issue composed of interlinked smaller issues, and address the entire integration process holistically from end-to-end, i.e. beginning with multiple data silos and ending with a linked dataset.

A comprehensive literature review identified appropriate theories, methodologies and techniques, evidenced by their use in prior research, each aimed at addressing an individual facet, i.e. domain understanding, semantic and syntactic interoperability, and data linkage strategies to create a linked dataset.

Moreover, the author aligned these informatics techniques sequentially making them 'fit together' to facilitate the creation of a conceptual framework, applied to a real-world scenario to overcome HIF and achieve data linkage and integration. An additional benefit of this process is that it does not rely on any particular software application.

7.4.2 Record linkage strategy

This is the first study to develop a hybrid algorithm in which the DRL strategy utilised the mode to deduplicate patient IDs and quasi-identifiers, based on the frequency of prior patient attendances. Additionally, the PRL strategy builds on the work of Fellegi & Sunter [209] by incorporating complementary techniques to enhance the linkage process. Techniques utilised include: Blocking [198]; Soundex codes [199]; Quasi-identifiers and Combination ID [202]; Split date of birth [198, 213]; Jaro-Winkler string comparator [199, 202, 229, 230]; and Relative frequency [199, 213].

The Hybrid strategy was validated by comparison with application of the NHS HES algorithm [206] and was found to link a significantly higher number of records. Additionally, these RL strategies were applied to the same dataset within the CDW using SQL script, as opposed to using a particular software application, which requires prior evaluation and often considered as 'black box' approaches.

7.4.3 Linked dataset analysis

This study bridges the gap between research and practice, by demonstrating that data linkage and integration of routine administrative data can be used to create and answer research questions that would not have been possible using unlinked data. To demonstrate the benefit of a linked dataset, this study appears to be the first to report on the following analyses for the study setting:

1. To calculate the transition time for patient transfer between UCC and ED services.
- Summary: using paper to the bridge the gap of fragmentation costs time and unnecessarily prolongs the pain or suffering experienced by patients waiting to receive treatment.

2. To distinguish the effect of utilising Linked and Unlinked datasets, using the ED 4 hour standard as a proxy measure.
- Summary: a lack of interoperability between CIS can be measured in units of Clinician time unnecessarily spent reviewing four-hour breaches that were flagged due to inaccurate time recording.

3. To compare the rate of emergency patient admissions pre/post UCC implementation at HH and CXH hospital sites.
- Summary: Analysis highlights the benefits of using linked datasets to undertake in-depth analysis of admission rates to inform the decision-making process and aid service evaluation.

In summary, disparate routine administrative data can be an invaluable source of information for health service research, and such data can gain completely new qualities via their integration, to reveal new insights and create new knowledge. Furthermore, this study has led to seven original peer-reviewed publications, as listed on page 23 of this thesis.

7.5 Limitations

7.5.1 Level of interoperability

The most important limitation to consider for any data integration study is the level of interoperability that exists between multiple departments and their respective CIS. As there is no 'one' overall CIS encompassing all hospital information [8] the patient EMR is fragmented amongst multi-vendor disparate departmental CIS, [12] each of which function as independent units. [14] These heterogeneous CIS do not 'talk' to one another and become isolated [15] due to a lack of interoperability [16] leading to the formation of data silos.

Furthermore, as CIS are vendor-locked, researchers are not permitted to make any changes. The alternative was for vendors to collaborate in order to enable interoperability; however, this would require time and effort on the part of vendors, and is not always possible due to inherent differences in technologies between legacy and modern CIS.

Ultimately, the limitation of a lack of interoperability influences and restricts the approach that researchers can take for integration studies. For example in this study, as there was no CIS interoperability, and data were held in silos, the most feasible approach was to undertake data extraction processes via the use of CDW technologies, that did not affect the underlying data.

7.5.2 Level of heterogeneity

The second most important limitation was the level of heterogeneity between the multi-vendor CIS, which arises due to the underlying specialty of the department in which the CIS is to be used, and the purpose of the CIS. Whereby, 'heterogeneity' incorporates the structure of the patient EMR and the data contained within, i.e. semantics (meaning), syntactic (formatting), structure (storage) [23] and associated data dictionaries.

The issue of heterogeneity will be characteristic of disparate data silos and is not something that researchers are able to predict, nor influence. Vendors may not discuss this with other vendors due to fear of divulging industry secrets. Rather, researchers must make appropriately informed decisions/approaches to tackle this issue.

For this study, one universal approach was taken applicable to all three CISs, first, liaising with the Database Managers for each CIS to obtain expert knowledge and have a point of reference. Second, undertaking system training to enable an understanding of nuances for each CIS. Third, utilising data dictionaries to obtain information, e.g. metadata, relational schemas and data structures. Four, referring to conceptualisation and standardisation techniques from the literature and NHS data dictionaries to identify commonalities amongst the disparate CIS. Five, careful planning and documentation of all steps to facilitate transparency.

Moreover, researchers undertaking such approaches will be able to make informed decisions and take appropriate measures to ensure the same techniques are applied to all datasets in order to eliminate bias, and ensure study data remain genuine.

7.5.3 NHS number data quality

Although hospital patient data are routinely collected within CIS for the management and treatment of patients, routinely collected data have the potential for re-use in many other tasks, e.g. research. [178, 179] However, there is no control over what data are available, presence of inconsistencies, amount of missing data and their overall quality. [177-179]

Therefore, for this study, a key limitation was the discovery that the NHS number DQ completeness was approximately 46% within the ED dataset. As this study used retrospective data, it was not possible to address this issue. This finding signified that the three datasets could not be readily linked using DRL strategies, and restricting all datasets to only those records in which the NHS matched would automatically exclude 54% of the total data, invalidating any subsequent conclusions drawn from such a small dataset, due to its representativeness and credibility.

This limitation, combined with the identification of heterogeneous CIS specific patient IDs, meant that neither the NHS number nor the patient IDs could be used for data linkage or integration purposes, leaving only patient quasi-identifiers, e.g. name, gender, date of birth.

Moreover, this limitation influenced the methodological approach for this study, whereby the literature identified PRL strategies to link datasets using quasi-identifiers. Furthermore, using the same datasets, the PRL strategy was compared against a Step Based DRL HES algorithm [206] to provide a means of statistically validating the PRL strategy.

7.5.4 Study setting

One of the main criticisms of case study as a research strategy relates to the generalisability and credibility of their findings and theoretical contributions to knowledge. [272, 291] However, the viewpoint, that findings from case study research cannot be applied beyond the cases studied, are mostly positivist criticisms based on statistical generalisation, which rely upon the use of representative random samples in order to extrapolate findings to a larger general population. [272, 292]

Case studies enable researchers to focus on one particular instance of a phenomenon [291] within the context of a real-life setting, to promote in-depth inquiry and understanding. [272, 290] Furthermore, the findings from just one case may identify a wealth of transferable information and knowledge applicable to similar settings; thus, shifting the focus of the case study from generalisability to transferability. [292]

This research uses a single case study strategy based within a large metropolitan London hospital to investigate the phenomenon of HIF and data silos. Moreover, the aim of this study was not to compare specific results from any specific analyses. Rather, it was to identify select informatics that enabled the integration and linkage of data silos and resulted in the development of a conceptual framework, which may be implemented within other hospital settings facing similar HIF scenarios.

7.6 Recommendations for future work

7.6.1 CDW linked dataset development

The CDW linked dataset created within this study has demonstrated its potential to generate new knowledge and offer new insights. The CDW architecture facilitates the automation of scripts to import, process, assess data quality and link data on a daily scheduled basis.

Importantly, CISs within ED are based on a timestamp EPR structure and do not readily facilitate investigating an individual patients record longitudinally. However, the record structure within the CDW linked dataset is structured similarly to a patient EHR, adding a new dimension to future analysis.

For example, ICHNT encompasses the ED services at St Mary's Hospital (SMH), whose data could also be included to permit comparison of EDs and patient profiles. Further work could involve extending the time frame of the underlying data to observe the effect of ED closure at the HH site in September 2014, and on the ED services at both CXH and HH sites. Additionally, an analysis of patients that repeat attend to ED can be undertaken, and the CDW can also be used for patient reconciliation centrally.

7.6.2 Data linkage

Data linkage and deduplication techniques are used within the healthcare sector for cleaning and compiling datasets for longitudinal or other epidemiological studies. [349] In recent years, research interest has increased, incorporating techniques originating from statistics, machine learning, information retrieval, and database research, all combined and applied to improve linkage quality and increase the performance of linkage and deduplication of very large data sets. [202, 349]

Although this study utilised complementary techniques to enhance the linkage process, further studies could, using a gold standard dataset, investigate the effect of each individual linkage strategy in a stepwise manner to quantify the effect of each individual technique. Additional techniques from the literature include Supervised Machine Learning, e.g. String distance, decision trees, nearest neighbour, or Unsupervised Machine Learning, e.g. clustering, graphical models or Hybrid approaches, SQL Extensions and Expert systems. [202, 349]

7.6.3 Primary care collaboration

This study focused specifically on the urgent and emergency care pathway, which incorporates the UCC, ED and IPC departments. The 'Partnership for Health' collaboration also provides an entirely separate primary care service which works in parallel with the GP-led UCC and is co-located alongside the EDs at both CXH and HH sites. However, this service works in complete isolation, operating independently from both the UCC and ED services, and was therefore not included in this research.

The primary care services operate between 0800 to 1830 Monday to Friday, and 0900 to 1300 on Saturdays and patients that have registered with this service need to book appointments to see a GP. However, there is some evidence to suggest that these registered patients often present to the UCC.

Further work could involve UCC and ED services engaging with the primary care service to cross-reference and identify those patients who, although registered, are at higher risk of attending or repeat presenting to the UCC service. Such an investigation would be possible by comparing the current linked dataset with the primary care dataset. This collaboration between services would help inform clinical and managerial stakeholders on the best course of action for patient management and service provision.



References

1. Dunn, H.L., *Record Linkage.* American Journal of Public Health and the Nations Health, 1946. **36**(12): p. 1412-1416.
2. Sood, H.S. and K. McNeil, *How is health information technology changing the way we deliver NHS hospital care?* Future Hospital Journal, 2017. **4**(2): p. 117-120.
3. Oxford Dictionaries. *Definition of Silo*. 2017 [cited 2017 04/11/2017]; Available from: https://en.oxforddictionaries.com/definition/silo.
4. Zipperer, L.A., *Patient safety : perspectives on evidence, information and knowledge transfer.* 2014.
5. Hersh, W., *Health care information technology: progress and barriers.* Jama, 2004. **292**(18): p. 2273-2274.
6. Vogel, L.H., *Management of Information in Health Care Organizations*, in *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, E.H. Shortliffe and J.J. Cimino, Editors. 2014, Springer London: London. p. 443-474.
7. Benson, T., *Why general practitioners use computers and hospital doctors do not—Part 2: scalability.* BMJ, 2002. **325**(7372): p. 1090-1093.
8. Hasselbring, W., *Federated integration of replicated information within hospitals.* International Journal on Digital Libraries, 1997. **1**(3): p. 192-208.
9. Wagner, G., et al., *Medical data management: A practical guide*. 2006: Springer Science & Business Media.
10. Sullivan, F. and J. Wyatt, *ABC of health informatics*. 2009: John Wiley & Sons.
11. Vest, J.R. and K. Simon, *Hospitals' adoption of intra-system information exchange is negatively associated with inter-system information exchange.* Journal of the American Medical Informatics Association, 2018. **25**(9): p. 1189-1196.
12. Chenhui, Z., D. Huilong, and L. Xudong. *An integration approach of healthcare information system*. in *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*. 2008. IEEE.
13. Zheng, R., et al. *Heterogeneous medical data share and integration on grid*. in *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*. 2008. IEEE.
14. Stead, W.W., et al., *Integration and BeyondLinking Information from Disparate Sources and into Workflow.* Journal of the American Medical Informatics Association, 2000. **7**(2): p. 135-145.
15. Kiourtis, A., et al., *Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching.* International Journal of Medical Informatics, 2019. **132**: p. 104002.
16. IEEE. *Interoperability definition*. 2017 [cited 2017 09/01/2017]; Available from: http://ieeexplore.ieee.org/xpls/dictionary.jsp?stdDict=browse_keyword&pageNumber=1&def_term=interoperability&def_id=&stdDictionary_tarid=&stdDictionary_tarn=&stdDictionary_scn=Aerospace+Electronics&nav=.
17. Grannis, S.J., et al., *Evaluating the effect of data standardization and validation on patient matching accuracy.* Journal of the American Medical Informatics Association, 2019. **26**(5): p. 447-456.
18. Coiera, E., *Clinical communication: a new informatics paradigm.* Proceedings of the AMIA Annual Fall Symposium, 1996: p. 17-21.
19. Coiera, E., *When Conversation Is Better Than Computation.* Journal of the American Medical Informatics Association, 2000. **7**(3): p. 277-286.
20. Parker, J. and E. Coiera, *Improving Clinical CommunicationA View from Psychology.* Journal of the American Medical Informatics Association, 2000. **7**(5): p. 453-461.

21. Coiera, E., *Information systems (Chapter 3)*, in *Guide to health informatics, Third edition* 2015, CRC Press. p. 23-34.
22. Coiera, E., *Information management systems (Chapter 9)*, in *Guide to health informatics, Third edition* 2015, CRC Press. p. 117-125.
23. Liyanage, H., P. Krause, and S. De Lusignan, *Using ontologies to improve semantic interoperability in health data.* J Innov Health Inform, 2015. **22**(2): p. 309-15.
24. Coiera, E., *Communicating (Chapter 4)*, in *Guide to health informatics, Third edition* 2015, CRC Press. p. 37-49.
25. NHS England. *Data and information*. 2017 [cited 2017 05/11/2017]; Available from: https://www.england.nhs.uk/ourwork/tsd/data-info/#.
26. Wyatt, J.C. and J.L.Y. Liu, *Basic concepts in medical informatics.* Journal of Epidemiology and Community Health, 2002. **56**(11): p. 808-812.
27. Haas, L. *Beauty and the beast: The theory and practice of information integration*. in *International Conference on Database Theory*. 2007. Springer.
28. Bygstad, B., O. Hanseth, and D. Truong Le, *From IT silos to integrated solutions. A study in e-health complexity.* 2015.
29. Xu, Y., et al., *Integration of medical applications: the 'mediator service'of the SynEx platform.* International Journal of Medical Informatics, 2000. **58**: p. 157-166.
30. Stange, K.C., *The Problem of Fragmentation and the Need for Integrative Solutions.* Annals of Family Medicine, 2009. **7**(2): p. 100-103.
31. Harrison, M.I., R. Koppel, and S. Bar-Lev, *Unintended consequences of information technologies in health care--an interactive sociotechnical analysis.* Journal of the American Medical Informatics Association : JAMIA, 2007. **14**(5): p. 542-549.
32. Booth, R., S. McBride, and M. Tietze, *Scientific and theoretical foundations for driving improvement.* Nursing Informatics for the Advanced Practice Nurse: Patient Safety, Quality, Outcomes, and Interprofessionalism, 2015.
33. Oliver, P., *Writing your thesis*. 2014.
34. Gnani, S., et al., *Evaluation of a general practitioner-led urgent care centre in an urban setting: description of service model and plan of analysis.* JRSM Short Rep, 2013. **4**(6): p. 2042533313486263.
35. Cowling, T.E., et al., *Attendances at Charing Cross and Hammersmith Hospitals' urgent care centres, 2009-12.* Bmj, 2013. **347**: p. f7035.
36. Cowling, T.E., et al., *Data from two London hospitals' urgent care centres, 2009-12.* BMJ: British Medical Journal, 2013. **347**(7935): p. 24-24.
37. Amiel, C., et al., *Reasons for attending an urban urgent care centre with minor illness: a questionnaire study.* Emerg Med J, 2014. **31**(e1): p. e71-5.
38. Gnani, S., et al., *Patterns of healthcare use among adolescents attending an urban general practitioner-led urgent care centre.* Emerg Med J, 2014. **31**(8): p. 630-6.
39. Cowling, T.E., et al., *Referral outcomes of attendances at general practitioner led urgent care centres in London, England: retrospective analysis of hospital administrative data.* Emerg Med J, 2015.
40. Gnani, S., et al., *Healthcare use among preschool children attending GP-led urgent care centres: a descriptive, observational study.* BMJ Open, 2016. **6**(6).
41. Alshaikh, F., et al., *Social network sites as a mode to collect health data: a systematic review.* J Med Internet Res, 2014. **16**(7): p. e171.
42. Creswell, J.W., *Research design : qualitative, quantitative, and mixed methods approaches*. 2014, Thousand Oaks, California: SAGE Publications.
43. O'Leary, Z., *The essential guide to doing research*. 2007, London [u.a.]: SAGE.
44. NHS. *The NHS in England*. 2016 [cited 2017 05/04/2017]; Available from: http://www.nhs.uk/NHSEngland/thenhs/about/Pages/overview.aspx.
45. Iacobucci, G., *NHS in 2017: the long arm of government.* BMJ, 2017. **356**.

46. NHS Providers. *The NHS provider sector*. 2017 [cited 2017 12/12/2017]; Available from: http://nhsproviders.org/topics/delivery-and-performance/the-nhs-provider-sector.
47. NHS. *The NHS in England - Structure*. 2016 [cited 2017 05/04/2017]; Available from: http://www.nhs.uk/NHSEngland/thenhs/about/Pages/nhsstructure.aspx.
48. NHS England. *Models of care*. 2017 [cited 2017 12/12/2017]; Available from: https://www.england.nhs.uk/new-care-models/about/.
49. NHS England. *Primary care home model*. 2017 [cited 2017 12/12/2017]; Available from: https://www.england.nhs.uk/new-care-models/pch/.
50. The Medic Portal. *Challenges Facing The NHS and Current NHS Issues*. 2017 [cited 2017 29/12/2017]; Available from: https://www.themedicportal.com/application-guide/the-nhs/challenges-facing-the-nhs/.
51. NHS London CCG. *Today's NHS - our current challenges*. 2017 [cited 2017 17/04/2017]; Available from: https://www.myhealth.london.nhs.uk/help/nhs-today.
52. ACEP. *Definition of Emergency Medicine*. 2017 [cited 2017 04/11/2017]; Available from: https://www.acep.org/Clinical---Practice-Management/Definition-of-Emergency-Medicine/#sm.0000cl8pk0wjueq2ult2owkbduxqw.
53. Fatovich, D.M., *Emergency medicine.* BMJ : British Medical Journal, 2002. **324**(7343): p. 958-962.
54. Fordyce, J., et al., *Errors in a busy emergency department.* Annals of Emergency Medicine, 2003. **42**(3): p. 324-333.
55. Institute of Medicine, *Building a 21st-Century Emergency and Trauma Care System*, in *Emergency Medical Services: At the Crossroads*. 2007, The National Academies Press: Washington, DC. p. 73-118.
56. NHS Digital. *NHS Business Definition: ED Type*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/attributes/a/acc/accident_and_emergency_department_type_de.asp.
57. NHS England. *A&E Attendances and Emergency Admissions Monthly Return Definitions*. 2017 [cited 2017 12/12/2017]; Available from: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2013/03/AE-Attendances-Emergency-Definitions-v2.0-Final.pdf.
58. Wyatt, J.P., *Oxford handbook of emergency medicine*. 2011, Oxford: Oxford University Press.
59. NHS Digital. *NHS Business Definition: A&E Attendance*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/a/accident_and_emergency_attendance_de.asp?shownav=1.
60. NHS. *The NHS in England - Urgent & Emergency Care*. 2015 [cited 2017 05/04/2017]; Available from: http://www.nhs.uk/NHSEngland/AboutNHSservices/Emergencyandurgentcareservices/Pages/AE.aspx.
61. NHS Digital. *NHS Business Definition: Admission Method*. 2016 [cited 2016 25/08/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/attributes/a/add/admission_method_de.asp.
62. House of Commons. *Urgent and emergency services*. 2013 [cited 2017 28/03/2017]; Available from: http://www.parliament.uk/business/committees/committees-a-z/commons-select/health-committee/news/13-07-23-urgemrepcs/.
63. The Royal College of Emergency Medicine. *Initial Assessment of Emergency Department Patients* 2017 03/06/2017]; Available from: http://www.rcem.ac.uk/docs/SDDC%20Intial%20Assessment%20(Feb%202017).pdf.
64. NHS Digital. *NHS Business Definition: A&E Initial Assessment Triage Category*. 2017 [cited 2017 25/07/2017]; Available from:

http://www.datadictionary.nhs.uk/data_dictionary/attributes/a/a_and_e_initial_assessment_triage_category_de.asp?shownav=1.

65. NHS Digital. *NHS Business Definition: Stream*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/attributes/a/a_and_e_stream_de.asp?shownav=1?query=%22stream%22&rank=100&shownav=1.
66. Carson, D., H. Clay, and R. Stern. *Primary Care and Emergency Departments*. 2010 [cited 2017 18/04/2017]; Available from: https://www.primarycarefoundation.co.uk/images/PrimaryCareFoundation/Downloading_Reports/Reports_and_Articles/Primary_Care_and_Emergency_Departments/Primary_Care_and_Emergency_Departments_RELEASE.pdf.
67. The Royal College of Emergency Medicine. *Emergency Department Clinical Quality Indicators: A CEM guide to implementation*. 2017 03/06/2017]; Available from: https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&cad=rja&uact=8&ved=0ahUKEwi8-Mv2rcDYAhWsDcAKHQqOB8AQFghHMAQ&url=https%3A%2F%2Fsecure.rcem.ac.uk%2Fcode%2Fdocument.asp%3FID%3D5832&usg=AOvVaw1RJ5VjAdoTQ-oNn6RaleKa.
68. Department of Health. *A&E clinical quality indicators: Implementation guidance and data definitions*. 2010 [cited 2017 18/04/2017]; Available from: http://webarchive.nationalarchives.gov.uk/20130107105354/http:/www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_122868.
69. Mason, S., et al., *Time Patients Spend in the Emergency Department: England's 4-Hour Rule—A Case of Hitting the Target but Missing the Point?* Annals of Emergency Medicine, 2011. **59**(5): p. 341-349.
70. National Audit Office. *Improving Emergency Care in England*. 2004 [cited 2017 18/04/2017]; Available from: https://www.nao.org.uk/wp-content/uploads/2004/10/03041075.pdf.
71. Nuffield Trust Health Foundation. *Quality Watch: A&E waiting times* 2017 [cited 2017 12/12/2017]; Available from: http://www.qualitywatch.org.uk/indicator/ae-waiting-times.
72. Office for National Statistics. *England population mid-year estimate*. 2018 [cited 2018 04/01/2018]; Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/timeseries/enpop/pop.
73. Fatovich, D.M. and R.L. Hirsch, *Entry overload, emergency department overcrowding, and ambulance bypass.* Emergency Medicine Journal, 2003. **20**(5): p. 406-409.
74. He, J., et al., *Demand for hospital emergency departments: a conceptual understanding.* World Journal of Emergency Medicine, 2011. **2**(4): p. 253-261.
75. Meggs, W.J., T. Czaplijski, and N. Benson, *Trends in Emergency Department Utilization, 1988-1997.* Academic Emergency Medicine, 1999. **6**(10): p. 1030-1035.
76. House of Commons. *Accident & Emergency Statistics: demand, performance & pressure*. 2017 [cited 2017 28/03/2017]; Available from: http://researchbriefings.files.parliament.uk/documents/SN06964/SN06964.pdf.
77. NHS England. *A&E Attendances and Emergency Admissions 2017-18*. 2017 [cited 2017 12/12/2017]; Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/ae-attendances-and-emergency-admissions-2017-18/.
78. National Audit Office. *Emergency admissions to hospital: managing the demand*. 2013 [cited 2017 18/04/2017]; Available from: https://www.nao.org.uk/wp-content/uploads/2013/10/10288-001-Emergency-admissions.pdf.
79. Simonet, D., *Cost reduction strategies for emergency services: insurance role, practice changes and patients accountability.* Health Care Analysis, 2009. **17**(1): p. 1-19.
80. Goodacre, S. and C. McCabe, *An introduction to economic evaluation.* Emergency medicine journal: EMJ, 2002. **19**(3): p. 198.

81. Derlet, R.W., *Overcrowding in emergency departments: increased demand and decreased capacity.* Annals of emergency medicine, 2002. **39**(4): p. 430-432.
82. Derlet, R.W. and J.R. Richards, *Overcrowding in the nation's emergency departments: complex causes and disturbing effects.* Annals of emergency medicine, 2000. **35**(1): p. 63-68.
83. Capewell, S., *The continuing rise in emergency admissions.* BMJ: British Medical Journal, 1996. **312**(7037): p. 991.
84. Murphy, A.W., *'Inappropriate'attenders at accident and emergency departments I: definition, incidence and reasons for attendance.* Family Practice, 1998. **15**(1): p. 23-32.
85. Moll van Charante, E.P., G. ter Riet, and P. Bindels, *Self-referrals to the A&E department during out-of-hours: patients' motives and characteristics.* Patient Educ Couns, 2008. **70**(2): p. 256-65.
86. Majeed, A., *Inversion of emergency pyramid. "I wanted a thorough checkup".* Bmj, 2003. **326**(7388): p. 553.
87. Uscher-Pines, L., et al., *Deciding to Visit the Emergency Department for Non-Urgent Conditions: A Systematic Review of the Literature.* The American journal of managed care, 2013. **19**(1): p. 47-59.
88. NHS England. *Transforming urgent and emergency care services in England*. 2015 [cited 2016 17/12/2016]; Available from: https://www.england.nhs.uk/wp-content/uploads/2015/06/trans-uec.pdf.
89. NHS England. *Next steps on the NHS Five Year Forward View*. 2017 [cited 2017 12/12/2017]; Available from: https://www.england.nhs.uk/publication/next-steps-on-the-nhs-five-year-forward-view/.
90. Boeke, A.J.P., et al., *Effectiveness of GPs in accident and emergency departments.* Br J Gen Pract, 2010. **60**(579): p. e378-e384.
91. Murphy, A.W., *'Inappropriate'attenders at accident and emergency departments II: health service responses.* Family practice, 1998. **15**(1): p. 33-37.
92. NHS. *Primary Care in emergency departments - a guide to good practice*. 2015 [cited 2018 21/01/2018]; Available from: http://www.nhsimas.nhs.uk/fileadmin/Files/IST/Emergency_care_conference_2014/Primary_Care_in_A_E_Guidance_Feb_2015.pdf.
93. Dale, J., et al., *Cost Effectiveness Of Treating Primary Care Patients In Accident And Emergency: A Comparison Between General Practitioners, Senior House Officers, And Registrars.* BMJ: British Medical Journal, 1996. **312**(7042): p. 1340-1344.
94. Dale, J., et al., *Primary care in the accident and emergency department: I. Prospective identification of patients.* BMJ, 1995. **311**(7002): p. 423-426.
95. Dale, J., et al., *Primary care in the accident and emergency department: II. comparison of general practitioners and hospital doctors.* BMJ, 1995. **311**(7002): p. 427-430.
96. Murphy, A.W., et al., *Randomised controlled trial of general practitioner versus usual medical care in an urban accident and emergency department: process, outcome, and comparative cost.* BMJ, 1996. **312**(7039): p. 1135-1142.
97. Oxford Dictionaries. *Definition of Information*. 2017 [cited 2017 05/11/2017]; Available from: https://en.oxforddictionaries.com/definition/information.
98. Wyatt, J.C. and F. Sullivan, *What is health information?* BMJ (Clinical research ed.), 2005. **331**(7516): p. 566-568.
99. Shortliffe, E.H. and J.J. Cimino, *Biomedical informatics : computer applications in health care and biomedicine.* 2014.
100. Ackoff, R.L., *From data to wisdom.* Journal of applied systems analysis, 1989. **16**(1): p. 3-9.
101. Coiera, E., *Information (Chapter 2)*, in *Guide to health informatics, Third edition* 2015, CRC Press. p. 13-22.
102. Bellinger, G., D. Castro, and A. Mills, *Data, information, knowledge, and wisdom.* 2004.

103. Shortliffe, E.H. and G.O. Barnett, *Biomedical Data: Their Acquisition, Storage, and Use*, in *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, E.H. Shortliffe and J.J. Cimino, Editors. 2014, Springer London: London. p. 39-66.
104. Medicine, I.o., *Technology and Communications* in *Hospital-Based Emergency Care: At the Breaking Point*. 2007, The National Academies Press: Washington, DC. p. 165-208.
105. Yasnoff, W.A., *Health Information Infrastructure*, in *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, E.H. Shortliffe and J.J. Cimino, Editors. 2014, Springer London: London. p. 423-441.
106. Oxford Dictionaries. *Definition of Communication*. 2017 [cited 2017 07/11/2017]; Available from: https://en.oxforddictionaries.com/definition/communication.
107. Wyatt, J.C. and P. Wright, *Design should help use of patients' data.* The Lancet, 1998. **352**(9137): p. 1375-1378.
108. Medicine, I.o., *To Err Is Human: Building a Safer Health System*, ed. L.T. Kohn, J.M. Corrigan, and M.S. Donaldson. 2000, Washington, DC: The National Academies Press. 312.
109. Coiera, E.W., et al., *Communication loads on clinical staff in the emergency department.* Med J Aust, 2002. **176**(9): p. 415-8.
110. Leape, L.L. and D.M. Berwick, *Safe health care: are we up to it? : We have to be.* BMJ : British Medical Journal, 2000. **320**(7237): p. 725-726.
111. Oxford Dictionaries. *Definition of Fragment*. 2017 [cited 2017 04/11/2017]; Available from: https://en.oxforddictionaries.com/definition/fragment.
112. Institute of Medicine, et al., *The computer-based patient record an essential technology for health care*. 1997, Washington, D.C.: National Academy Press.
113. Coiera, E., *The electronic health record (Chapter 10)*, in *Guide to health informatics, Third edition* 2015, CRC Press. p. 127-144.
114. Ledbetter, C.S. and M.W. Morgan, *Toward best practice: leveraging the electronic patient record as a clinical data warehouse.* Journal of Healthcare Information Management, 2001. **15**(2): p. 119-132.
115. Hersh, W.R., *Medical informatics: Improving health care through information.* JAMA, 2002. **288**(16): p. 1955-1958.
116. Grimson, J., W. Grimson, and W. Hasselbring, *The SI challenge in health care.* Communications of the ACM, 2000. **43**(6): p. 48-55.
117. Sullivan, F. and J.C. Wyatt, *Why is this patient here today?* BMJ : British Medical Journal, 2005. **331**(7518): p. 678-680.
118. Coiera, E., *Structuring (Chapter 5)*, in *Guide to health informatics, Third edition* 2015, CRC Press. p. 51-62.
119. Currie, W. and D. Finnegan, *Integrating Healthcare with Information and Communications Technology.* 2016.
120. Kent, W., *The breakdown of the information model in multi-database systems.* ACM SIGMOD Record, 1991. **20**(4): p. 10-15.
121. Codd E. F, *A relational model of data for large shared data banks.* Commun. ACM, 1970. **13**(6): p. 377-387.
122. Connolly, T.M. and C.E. Begg, *Database systems : a practical approach to design, implementation, and management*. 2014.
123. Coiera, E., *Guide to health informatics*. Third Edition ed. 2015: CRC Press.
124. Friedman, C., et al., *A Generalized Relational Schema for an Integrated Clinical Patient Database.* Proceedings of the Annual Symposium on Computer Application in Medical Care, 1990: p. 335-339.
125. McBride, S. and M. Tietze, *Nursing informatics for the advanced practice nurse : patient safety, quality, outcomes, and interprofessionalism.* 2016.

126. Koppel, R. and C.U. Lehmann, *Implications of an emerging EHR monoculture for hospitals and healthcare systems.* Journal of the American Medical Informatics Association, 2014. **22**(2): p. 465-471.
127. Nugawela, S. and T. Sahama. *Barriers to the adoption of health information technology*. in *Proceedings of International Conference on Innovation and Management (IAM), Kuala Lumpur, Malaysia*. 2011.
128. Miller, A.R. and C. Tucker, *Health information exchange, system size and information silos.* Journal of Health Economics, 2014. **33**: p. 28-42.
129. Marsolo, K., *Informatics and operations—let's get integrated.* Journal of the American Medical Informatics Association, 2012. **20**(1): p. 122-124.
130. NHS England. *Interoperability Handbook*. 2017 [cited 2017 09/01/2017]; Available from: https://www.england.nhs.uk/publication/interoperabilty-handbk/.
131. Commission of the European Communities. *Commission Recommendation on cross-border interoperability of electronic health record systems*. 2008 [cited 2018 04/12/2018]; Available from: https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-cross-border-interoperability-electronic-health-record-systems.
132. Benson, T. and G. Grieve, *Principles of health interoperability : SNOMED CT, HL7 and FHIR.* 2016.
133. Gagnon, M. *Ontology-based integration of data sources*. in *2007 10th International Conference on Information Fusion*. 2007. IEEE.
134. Department of Health. *Making IT work: harnessing the power of health information technology to improve care in England* 2016 [cited 2017 16/11/2017]; Available from: https://www.gov.uk/government/publications/using-information-technology-to-improve-the-nhs/making-it-work-harnessing-the-power-of-health-information-technology-to-improve-care-in-england.
135. Warner, H.R., *Medical informatics: a real discipline?* Journal of the American Medical Informatics Association, 1995. **2**(4): p. 207.
136. Raghupathi, W. and J. Tan, *Strategic IT applications in health care.* Commun. ACM, 2002. **45**(12): p. 56-61.
137. Das, S.K., *A scheme for classifying integration types in CIM.* International Journal of Computer Integrated Manufacturing, 1992. **5**(1): p. 10-17.
138. Wainwright, D. and T. Waring, *Three domains for implementing integrated information systems: redressing the balance between technology, strategic and organisational analysis.* International Journal of Information Management, 2004. **24**(4): p. 329-346.
139. Prather, J.C., et al. *Medical data mining: knowledge discovery in a clinical data warehouse*. in *Proceedings of the AMIA annual fall symposium*. 1997. American Medical Informatics Association.
140. Bettencourt-Silva, J., et al., *On creating a patient-centric database from multiple Hospital Information Systems.* Methods Inf Med, 2012. **51**(3): p. 210-20.
141. Greenes, R.A., *Clinical decision support : the road to broad adoption.* 2014.
142. U.S. National Library of Medicine. *Medical Subject Headings 2020*. 2016 [cited 2019 14/10/2019]; Available from: https://meshb.nlm.nih.gov/search.
143. Coletti, M.H. and H.L. Bleich, *Medical Subject Headings Used to Search the Biomedical Literature.* Journal of the American Medical Informatics Association, 2001. **8**(4): p. 317-323.
144. Rector, A. and D. Sottara, *Chapter 20 - Formal Representations and Semantic Web Technologies*, in *Clinical Decision Support (Second Edition)*, R.A. Greenes, Editor. 2014, Academic Press: Oxford. p. 551-598.
145. Butterfield, A., G.E. Ngondi, and A. Kerr, *A dictionary of computer science.* 2016.
146. Pérez-Rey, D., et al., *ONTOFUSION: Ontology-based integration of genomic and clinical databases.* Computers in Biology and Medicine, 2006. **36**(7): p. 712-730.

147. de Lusignan, S., et al., *Defining datasets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: an ontology-driven approach.* Journal of Innovation in Health Informatics, 2011. **19**(3): p. 127-134.
148. Gruber, T.R., *Toward principles for the design of ontologies used for knowledge sharing?* International Journal of Human-Computer Studies, 1995. **43**(5): p. 907-928.
149. Jakus, G., et al., *Concepts, Ontologies, and Knowledge Representation.* 2013.
150. Fonseca, F., *The double role of ontologies in information science research.* Journal of the American Society for Information Science and Technology, 2007. **58**(6): p. 786-793.
151. Guarino, N., *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*. Vol. 46. 1998: IOS press.
152. Rubin, D.L., et al., *National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge.* Omics, 2006. **10**(2): p. 185-98.
153. Valaski, J., A. Malucelli, and S. Reinehr, *Ontologies application in organizational learning: A literature review.* Expert Systems with Applications, 2012. **39**(8): p. 7555-7561.
154. Liaw, S.-T., et al., *Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease registers.* Journal of Biomedical Informatics, 2014. **52**: p. 364-372.
155. Liyanage, H., et al., *The Evidence-base for Using Ontologies and Semantic Integration Methodologies to Support Integrated Chronic Disease Management in Primary and Ambulatory Care: Realist Review.* Yearbook of Medical Informatics, 2013. **22**(01): p. 147-154.
156. Dixon, B.E., D.J. Vreeman, and S.J. Grannis, *The long road to semantic interoperability in support of public health: Experiences from two states.* Journal of Biomedical Informatics, 2014. **49**: p. 3-8.
157. Wache, H., et al. *Ontology-Based Integration of Information-A Survey of Existing Approaches*. in *OIS@ IJCAI*. 2001.
158. Goh, C.H., *Representing and reasoning about semantic conflicts in heterogeneous information systems*. 1997, Massachusetts Institute of Technology.
159. Greenbaum, N.R., et al., *Improving documentation of presenting problems in the emergency department using a domain-specific ontology and machine learning-driven user interfaces.* International Journal of Medical Informatics, 2019. **132**: p. 103981.
160. HL7 FHIR Foundation. *HL7 FHIR Release 4: Welcome*. 2019 [cited 2019 04/12/2019]; Available from: http://www.hl7.org/implement/standards/fhir/index.html.
161. Kilintzis, V., et al., *Supporting integrated care with a flexible data management framework built upon Linked Data, HL7 FHIR and ontologies.* Journal of Biomedical Informatics, 2019. **94**: p. 103179.
162. HL7 International. *HL7 Version 3 Product Suite*. 2019 [cited 2019 04/12/2019]; Available from: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=186.
163. De Lusignan, S., et al., *Key concepts to assess the readiness of data for International research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation.* Yearbook of medical informatics, 2011. **20**(01): p. 112-120.
164. HL7 FHIR Foundation. *HL7 FHIR Release 4: Resource Patient - Detailed Descriptions*. 2019 [cited 2019 04/12/2019]; Available from: http://www.hl7.org/implement/standards/fhir/patient-definitions.html.
165. NHS Digital. *NHS Digital: Fast Healthcare Interoperability Resources (FHIR)*. 2019 [cited 2019 04/12/2019]; Available from: https://digital.nhs.uk/services/fhir-apis.
166. Azevedo, A., *Data mining and knowledge discovery in databases*, in *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*. 2019, IGI Global. p. 502-514.
167. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *From data mining to knowledge discovery in databases.* AI magazine, 1996. **17**(3): p. 37.

168. Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth. *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. in *KDD*. 1996.
169. Burton, P.R., et al., *Data Safe Havens in health research and healthcare.* Bioinformatics, 2015. **31**(20): p. 3241-8.
170. Duncan, J., et al., *Building an Ontology for Identity Resolution in Healthcare and Public Health.* Online journal of public health informatics, 2015. **7**(2): p. e219-e219.
171. Reeves, L., *A manager's guide to data warehousing.* 2013.
172. Batini, C. and M. Scannapieco, *Data Quality : Concepts, Methodologies and Techniques*. 2006.
173. Inmon, W.H., *Building the data warehouse*. 2002, New York: J. Wiley.
174. Kimball, R. and M. Ross, *The data warehouse toolkit : the complete guide to dimensional modeling*. 2002, New York: Wiley.
175. Newcombe, H.B., *Handbook of record linkage: methods for health and statistical studies, administration, and business*. 1988: Oxford University Press, Inc.
176. RCP. *Hospital activity data, a guide for Clinicians* 2007 [cited 2018 03/02/2018]; Available from: http://content.digital.nhs.uk/media/1595/Hospital-Activity-Data-A-guide-for-clinicians/pdf/Hospital_Activity_Data_-_A_guide_for_clinicians.pdf.
177. Christen, P., *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. 2012: Springer Science & Business Media.
178. Hersh, W.R., et al., *Caveats for the use of operational electronic health record data in comparative effectiveness research.* Medical care, 2013. **51**(8 0 3): p. S30.
179. Peacock, J. and P.J. Peacock, *Oxford handbook of medical statistics.* 2011.
180. PHAST Public Health Action Support Team CIC. *Introduction to routine data*. 2017 [cited 2017 30/07/2017]; Available from: https://www.healthknowledge.org.uk/e-learning/health-information/population-health-practitioners/introduction-routine-data.
181. Davoudi, S., et al., *Data Quality Management Model (2015 Update).* Journal of AHIMA, 2015. **86**(10): p. expanded web version.
182. Iyamu, T. and S. Mgudlwa, *Transformation of healthcare big data through the lens of actor network theory.* International Journal of Healthcare Management, 2018. **11**(3): p. 182-192.
183. The International Standards Organization (ISO). *ISO Online Browsing Platform: Data Quality*. 2019 [cited 2019 04/12/2019]; Available from: https://www.iso.org/obp/ui#home.
184. de Lusignan, S., *The optimum granularity for coding diagnostic data in primary care: report of a workshop of the EFMI Primary Care Informatics Working Group at MIE 2005.* Journal of Innovation in Health Informatics, 2006. **14**(2): p. 133-137.
185. Weidema, B.P. and M.S. Wesnæs, *Data quality management for life cycle inventories—an example of using data quality indicators.* Journal of Cleaner Production, 1996. **4**(3): p. 167-174.
186. Liaw, S.T., et al., *Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature.* International Journal of Medical Informatics, 2013. **82**(1): p. 10-24.
187. Wang, R.Y. and D.M. Strong, *Beyond accuracy: What data quality means to data consumers.* Journal of management information systems, 1996. **12**(4): p. 5-33.
188. Olson, J.E., *Data quality : the accuracy dimension*. 2008, Amsterdam [u.a.: Morgan Kaufmann.
189. DAMA UK. *The six primary dimensions for data quality assessment*. 2013 [cited 2017 02/11/2017]; Available from: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf.
190. CDC. *The Six Dimensions of Data Quality Assessment*. 2016 [cited 2017 02/11/2017]; Available from: https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf.

191. NHS Digital. *NHS Digital: Data quality*. 2019 [cited 2019 04/12/2019]; Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/data-quality.
192. Liaw, S.-T., et al., *An integrated organisation-wide data quality management and information governance framework: theoretical underpinnings.* Journal of Innovation in Health Informatics, 2014. **21**(4): p. 199-206.
193. NHS Digital. *NHS Business Definition: Nhs Number*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/attributes/n/nhs/nhs_number_de.asp?shownav=1?query=%22nhs+number%22&rank=100&shownav=1.
194. NHS Digital. *NHS Number Validation*. 2016 [cited 2016 24/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/attributes/n/nhs/nhs_number_de.asp?shownav=1.
195. Liaw, S.-T., et al., *Data extraction from electronic health records-existing tools may be unreliable and potentially unsafe.* Australian Family Physician, 2013. **42**(11): p. 820.
196. Liaw, S.-T., et al. *Data quality and fitness for purpose of routinely collected data–a general practice case study from an electronic Practice-Based Research Network (ePBRN)*. in *AMIA Annual Symposium Proceedings*. 2011. American Medical Informatics Association.
197. Herzog, T.H., F. Scheuren, and W.E. Winkler, *Record linkage.* Wiley Interdisciplinary Reviews: Computational Statistics, 2010. **2**(5): p. 535-543.
198. Blakely, T. and C. Salmond, *Probabilistic record linkage and a method to calculate the positive predictive value.* International journal of epidemiology, 2002. **31**(6): p. 1246-1252.
199. Sayers, A., et al., *Probabilistic record linkage.* International Journal of Epidemiology, 2015.
200. Jutte, D.P., L.L. Roos, and M.D. Brownell, *Administrative record linkage as a tool for public health research.* Annual review of public health, 2011. **32**: p. 91-108.
201. Newcombe, H.B., et al., *Automatic linkage of vital records.* Science, 1959. **130**(3381): p. 954-9.
202. Shlomo, N., *Overview of Data Linkage Methods for Policy Design and Evaluation*, in *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design*, N. Crato and P. Paruolo, Editors. 2019, Springer International Publishing: Cham. p. 47-65.
203. Karmel, R., et al., *Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study.* BMC health services research, 2010. **10**(1): p. 41.
204. Harron, K., et al., *Linking Data for Mothers and Babies in De-Identified Electronic Health Data.* PloS one, 2016. **11**(10): p. e0164667.
205. Pacheco, A.G., et al., *Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in Brazil.* American journal of epidemiology, 2008. **168**(11): p. 1326-1332.
206. NHS Digital. *HES-DID Linkage Report*. 2016 [cited 2016 30/12/2016]; Available from: https://www.gov.uk/government/statistics/hes-did-data-linkage-report-provisional-summary-statistics-april-to-may-2016.
207. NHS England. *NHS Hospital data and datasets: a consultation*. 2013 [cited 2016 30/12/2016]; Available from: https://www.england.nhs.uk/wp-content/uploads/2013/07/hosp-data-consult.pdf.
208. Alvey, W. and B. Jamerson. *Record linkage techniques–1997*. in *Proceedings of an International Workshop and Exposition. March*. 1997.
209. Fellegi, I.P. and A.B. Sunter, *A Theory for Record Linkage.* Journal of the American Statistical Association, 1969. **64**(328): p. 1183-1210.
210. Michelson, M. and C.A. Knoblock. *Learning blocking schemes for record linkage*. in *AAAI*. 2006.

211. Waien, S.A., *Linking large administrative databases: a method for conducting emergency medical services cohort studies using existing data.* Academic emergency medicine, 1997. **4**(11): p. 1087-1095.
212. Campbell, K.M., D. Deck, and A. Krupski, *Record linkage software in the public domain: a comparison of Link Plus, The Link King, and abasic'deterministic algorithm.* Health Informatics Journal, 2008. **14**(1): p. 5-15.
213. Zhu, V.J., et al., *An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling.* J Am Med Inform Assoc, 2009. **16**(5): p. 738-45.
214. Cameron, C.M., et al., *Population health and clinical data linkage: The importance of a population registry.* Australian and New Zealand journal of public health, 2007. **31**(5): p. 459-463.
215. Boyd, J.H., et al., *Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia.* BMC health services research, 2015. **15**(1): p. 312.
216. Dean, J.M., et al., *Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services.* Annals of emergency medicine, 2001. **37**(6): p. 616-626.
217. Downing, A., R. Wilson, and M. Cooke, *Linkage of ambulance service and accident and emergency department data: a study of assault patients in the west midlands region of the UK.* Injury, 2005. **36**(6): p. 738-744.
218. DuVall, S.L., R.A. Kerber, and A. Thomas, *Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators.* Journal of biomedical informatics, 2010. **43**(1): p. 24-30.
219. Hermansen, S.W., M.F. Leitzmann, and A. Schatzkin, *The impact on National Death Index ascertainment of limiting submissions to Social Security Administration Death Master File matches in epidemiologic studies of mortality.* American journal of epidemiology, 2009: p. kwn404.
220. Kabudula, C.W., et al., *The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa.* BMC medical research methodology, 2014. **14**(1): p. 71.
221. Li, B., et al., *Assessing record linkage between health care and Vital Statistics databases using deterministic methods.* BMC health services research, 2006. **6**(1): p. 48.
222. Li, X., et al. *An empiric weight computation for record linkage using linearly combined fields' similarity scores*. in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. 2014. IEEE.
223. Méray, N., et al., *Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number.* Journal of clinical epidemiology, 2007. **60**(9): p. 883. e1-883. e11.
224. Morgan, A.S., et al., *Investigating increased admissions to neonatal intensive care in England between 1995 and 2006: data linkage study using Hospital Episode Statistics.* BMC medical research methodology, 2016. **16**(1): p. 57.
225. Poon, S.K., et al. *An Ensemble Approach for Record Matching in Data Linkage*. in *Digital Health Innovation for Consumers, Clinicians, Connectivity and Community: Selected Papers from the 24th Australian National Health Informatics Conference (HIC 2016)*. 2016. IOS Press.
226. National Archives. *Soundex System*. 2007 [cited 2015 25/04/2015]; Available from: https://www.archives.gov/research/census/soundex.html.
227. Winkler, W.E., *Matching and record linkage.* Business survey methods, 1995. **1**: p. 355-384.
228. Campbell, K.M., *Impact of record-linkage methodology on performance indicators and multivariate relationships.* Journal of substance abuse treatment, 2009. **36**(1): p. 110-117.
229. Jaro, M.A., *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.* Journal of the American Statistical Association, 1989. **84**(406): p. 414-420.

230. Winkler, W.E., *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.* 1990.
231. Pollock, J.J. and A. Zamora, *Automatic spelling correction in scientific and scholarly text.* Communications of the ACM, 1984. **27**(4): p. 358-368.
232. SAS Institute. *SAS*. 2013 [cited 2013 05/07/2013]; Available from: https://www.sas.com/en_gb/software/stat.html.
233. Microsoft. *SQL Server 2014*. 2013 [cited 2013 05/07/2013]; Available from: https://www.microsoft.com/en-gb/sql-server/.
234. Microsoft. *Microsoft Excel*. 2013 [cited 2013 05/07/2013]; Available from: https://products.office.com/en-gb/excel.
235. Statacorp. *Stata 14: Statistical analysis & statistical software*. 01/03/2016]; Available from: http://www.stata.com/.
236. The R Foundation. *The R Project for Statistical Computing*. 2013 [cited 2013 05/07/2013]; Available from: https://www.r-project.org/about.html.
237. Python Software Foundation. *Python Programming Language*. 2013 [cited 2013 05/07/2013]; Available from: https://www.python.org/.
238. Harron, K., et al., *Opening the black box of record linkage.* Journal of Epidemiology and Community Health, 2012. **66**(12): p. 1198-1198.
239. NHS North West London Collaborative. *Whole Systems Integrated Care (WSIC) Dashboards and Information Sharing*. 2020 [cited 2020 08/01/2020]; Available from: https://www.healthiernorthwestlondon.nhs.uk/news-resources/information-sharing.
240. Wilkins, J. *Why more joined-up data sources are crucial for the future of healthcare*. 2020 [cited 2020 08/01/2020]; Available from: https://imperialcollegehealthpartners.com/why-more-joined-up-data-sources-are-crucial-for-the-future-of-healthcare/.
241. CPRD. *Clinical Practice Research Datalink*. 2016 [cited 2016 25/07/2016]; Available from: https://www.cprd.com/.
242. EuroReach Project. *Data Linkage*. 2013 [cited 2016 25/07/2016]; Available from: http://www.healthdatanavigator.eu/data-management/data-linkage.
243. Padmanabhan, S., et al., *Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications.* European Journal of Epidemiology, 2019. **34**(1): p. 91-99.
244. Gill, L., et al., *Computerised linking of medical records: methodological guidelines.* Journal of Epidemiology and Community Health, 1993. **47**(4): p. 316-319.
245. Acheson, E.D. and J.G. Evans, *The Oxford Record Linkage Study: A Review of the Method with some Preliminary Results.* Proceedings of the Royal Society of Medicine, 1964. **57**(4): p. 269-274.
246. Holman, C.A.J., et al., *Population-based linkage of health records in Western Australia: development of a health services research linked database.* Australian and New Zealand journal of public health, 1999. **23**(5): p. 453-459.
247. Lyons, R.A., et al., *The SAIL databank: linking multiple health and social care datasets.* BMC medical informatics and decision making, 2009. **9**(1): p. 3.
248. Ford, D.V., et al., *The SAIL Databank: building a national architecture for e-health research and evaluation.* BMC health services research, 2009. **9**(1): p. 157.
249. LoBiondo-Wood, G. and J. Haber, *Nursing research : methods and critical appraisal for evidence-based practice*. 2014.
250. Grove, S.K., et al., *Understanding nursing research : building an evidence-based practice*. 2015.
251. Cohen, L., L. Manion, and K. Morrison, *Research methods in education*. 2018.
252. Teater, B., *An introduction to applying social work theories and methods.* 2014.
253. Andreae, D., *General systems theory: Contributions to social work theory and practice.* Social work treatment: Interlocking theoretical approaches, 2011. **5**: p. 242-254.

254. Nelson, R., *Major theories supporting health care informatics.* Health care informatics: An interdisciplinary approach, 2002: p. 3-27.
255. Whitworth, B., et al., *Expanding the criteria for evaluating socio-technical software.* IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2008. **38**(4): p. 777-790.
256. Whitworth, B. and M. Zaic, *The WOSP model: Balanced information system design and evaluation.* Communications of the Association for Information Systems, 2003. **12**(1): p. 17.
257. Berg, M., *Patient care information systems and health care work: a sociotechnical approach.* International Journal of Medical Informatics, 1999. **55**(2): p. 87-101.
258. Whitworth, B., *The social requirements of technical systems*, in *Virtual Communities: Concepts, Methodologies, Tools and Applications*. 2011, IGI Global. p. 1461-1481.
259. Walker, G.H., et al., *A review of sociotechnical systems theory: a classic concept for new command and control paradigms.* Theoretical Issues in Ergonomics Science, 2008. **9**(6): p. 479-499.
260. Gaskin, G.L., et al., *Sociotechnical Challenges of Developing an Interoperable Personal Health Record.* Applied clinical informatics, 2011. **2**(04): p. 406-419.
261. Adaba, G.B. and Y. Kebebew, *Improving a health information system for real-time data entries: An action research project using socio-technical systems theory.* Informatics for Health and Social Care, 2018. **43**(2): p. 159-171.
262. Al-Hakim, L., *Adapted Lean thinking for emergency departments: Information quality perspective*, in *Lean Thinking for Healthcare*. 2014, Springer. p. 97-111.
263. Coiera, E., *Four rules for the reinvention of health care.* BMJ, 2004. **328**(7449): p. 1197-1199.
264. Sittig, D.F. and H. Singh, *A new sociotechnical model for studying health information technology in complex adaptive healthcare systems.* Quality & safety in health care, 2010. **19 Suppl 3**(Suppl 3): p. i68-i74.
265. Feufel, M.A., F.E. Robinson, and V.L. Shalin, *The impact of medical record technologies on collaboration in emergency medicine.* International Journal of Medical Informatics, 2011. **80**(8): p. e85-e95.
266. Rohatgi, M. and G. Friedman. *A structured approach for assessing & analyzing technical & nontechnical interoperability in socio-technical systems*. in *2010 IEEE International Systems Conference*. 2010. IEEE.
267. Muhammad, I., S.Y. Teoh, and N. Wickramasinghe, *Trying to streamline healthcare delivery in Australia via the Personally Controlled Electronic Health Record (PCEHR)*, in *Lean thinking for healthcare*. 2014, Springer. p. 187-206.
268. Carayon, P., et al., *Advancing a sociotechnical systems approach to workplace safety – developing the conceptual framework.* Ergonomics, 2015. **58**(4): p. 548-564.
269. Gilson, L., et al., *Health policy and systems research : a methodology reader.* 2012.
270. Miles, M.B., A.M. Huberman, and J. Saldaña, *Qualitative data analysis : a methods sourcebook*. 2014.
271. Maxwell, J.A., *Qualitative research design an interactive approach*. 2013, Thousand Oaks (Calif.): SAGE Publications.
272. Saunders, M. and P. Tosey. *The Layers of Research Design*. 2013 [cited 2019 22/04/2019]; Available from: https://anlp.org/files/research-onion-layers_42_357.pdf.
273. Crotty, M., *The foundations of social research : meaning and perspective in the research process*. 1998, London; Thousand Oaks, Calif.: Sage Publications.
274. Johnson, B., *Educational research : quantitative, qualitative, and mixed approaches*. 2014.
275. Mertens, D.M., *Research and evaluation in education and psychology : integrating diversity with quantitative, qualitative, and mixed methods*. 2010, Los Angeles: Sage.
276. Bryman, A., *Social research methods*. 2012, Oxford; New York: Oxford University Press.
277. Snieder, R. and K.L. Larner, *The art of being a scientist : a guide for graduate students and their mentors*. 2009, Cambridge; New York: Cambridge University Press.

278. Neuman, W.L., *Workbook for Neumann Social research methods : qualitative and quantitative approaches*. 2014: Pearson.
279. Gray, D.E., *Doing research in the real world*. 2018.
280. Saunders, M.N.K., P. Lewis, and A. Thornhill, *Research methods for business students*. 2016.
281. Mertler, C.A., *Introduction to educational research*. 2019.
282. Leedy, P.E. and J.E. Ormrod, *Practical research : planning and design*. 2015, Boston: Pearson.
283. Warren-Gash, C. *Linking and sharing routine health data for research*. 2017 [cited 2019 25/11/2019]; Available from: https://www.phgfoundation.org/report/linking-and-sharing-routine-health-data-for-research.
284. Zozus, M., et al. *Acquiring and using electronic health record data*. 2015 [cited 2019 05/12/2019]; Available from: https://rethinkingclinicaltrials.org/resources/acquiring-and-using-electronic-health-record-data/.
285. Cowie, M.R., et al., *Electronic health records to facilitate clinical research.* Clinical research in cardiology : official journal of the German Cardiac Society, 2017. **106**(1): p. 1-9.
286. Hogan, H., et al., *Observational study to determine the utility of hospital administrative data to support case finding of English patients at higher risk of severe healthcare-related harm.* BMJ Open, 2019. **9**(6): p. e025372.
287. Polling, C., et al., *Using routine clinical and administrative data to produce a dataset of attendances at Emergency Departments following self-harm.* BMC emergency medicine, 2015. **15**: p. 15-15.
288. NHS Health Research Authority. *Determine whether your study is research*. 2016 [cited 2016 05/09/2016]; Available from: http://www.hra.nhs.uk/research-community/before-you-apply/determine-whether-your-study-is-research/.
289. ICHNT. *Imperial College Healthcare NHS Trust*. 2017 [cited 2017 13/03/2017]; Available from: https://www.imperial.nhs.uk/.
290. Yin, R.K. and D.T. Campbell, *Case study research and applications : design and methods*. 2018.
291. Denscombe, M., *The good research guide : for small-scale social research projects*. 2017.
292. Frey, B.B., *The Sage encyclopedia of educational research, measurement, and evaluation.* 2018.
293. Reynolds, M., et al., *Using discrete event simulation to design a more efficient hospital pharmacy for outpatients.* Health care management science, 2011. **14**(3): p. 223-236.
294. Babbie, E.R., *The basics of social research*. 2017.
295. Hammersmith & Fulham Council. *2011 Census Statistics: Hammersmith & Fulham*. 2011 [cited 2017 18/04/2017]; Available from: https://www.lbhf.gov.uk/sites/default/files/section_attachments/2011_census_statistics_hammersmith_and_fulham_briefing.pdf.
296. Hammersmith & Fulham Council. *Hammersmith & Fulham Borough Profile 2018*. 2018 [cited 2019 18/04/2019]; Available from: https://www.lbhf.gov.uk/sites/default/files/section_attachments/borough-profile-2018.pdf.
297. Department for Communities and Local Government. *English indices of deprivation 2010*. 24 March 2011 13/08/2014]; Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010.
298. LBHF. *London Borough Hammersmith & Fulham Hospitals and GPs*. 2018 [cited 2018 02/01/2018]; Available from: https://www.lbhf.gov.uk/health-and-care/health-and-disabilities/hospitals-and-gps.
299. ICHNT. *Imperial College Healthcare NHS Trust - Formation*. 2007 [cited 2017 13/03/2017]; Available from: http://www3.imperial.ac.uk/newsandeventspggrp/imperialcollege/newssummary/news_29-8-2007-10-13-58.

300. ICHNT. *Imperial College Healthcare NHS Trust: UCC*. 2017 [cited 2017 13/03/2017]; Available from: https://www.imperial.nhs.uk/our-services/emergency-care/urgent-care-centres.
301. HFCfH. *Hammersmith & Fulham Centres for Health* 2017 [cited 2017 13/03/2017]; Available from: http://www.hammersmithandfulhamcentresforhealth.nhs.uk/.
302. LCW. *London Central & West Unscheduled Care Collaborative*. 2017 [cited 2017 13/03/2017]; Available from: http://www.lcwucc.com/.
303. The Royal College of Emergency Medicine. *Medical and Practitioner Staffing in Emergency Departments*. 2015 09/09/2019]; Available from: https://www.rcem.ac.uk/docs/Workforce/RCEM%20Medical%20and%20Practitioner%20Staffing%20in%20EDs.pdf.
304. Álvarez, L.G., et al., *Data linkage between existing healthcare databases to support hospital epidemiology.* Journal of Hospital Infection, 2011. **79**(3): p. 231-235.
305. Advanced-Healthcare. *Adastra Clinical Patient Management System*. 2015 [cited 2015 08/07/2015]; Available from: http://www.advancedcomputersoftware.com/ahc/products/adastra-patient-management-system.php.
306. EMIS Health. *Emergency Care IT system - Symphony*. 2015 [cited 2015 01/05/2015]; Available from: https://www.emishealth.com/products/symphony/.
307. MAXIMS, I. *IMS MAXIMS Patient Administration System - ICHIS*. 2010 [cited 2015 01/05/2015]; Available from: http://www.imsmaxims.com/imperial-college-healthcare-nhs-trust-supported-by-ims-maxims-patient-administration-system/.
308. Kumarapeli, P., et al., *Using Unified Modelling Language (UML) as a process-modelling technique for clinical-research process improvement.* Medical Informatics and the Internet in Medicine, 2007. **32**(1): p. 51-64.
309. Dennis, A., et al., *System analysis & design : an object-oriented approach with UML*. 2015.
310. Fakhroutdinov, K. *UML Activity Diagram Controls*. 2019 [cited 2019 22/10/2019]; Available from: https://www.uml-diagrams.org/.
311. NHS Digital. *Person Diagram Overview*. 2019 [cited 2019 04/12/2019]; Available from: https://www.datadictionary.nhs.uk/data_dictionary/diagrams/diagrams/person_diagram_fr.asp?shownav=1.
312. NHS Digital. *NHS Business Definition: Hospital Provider Spell*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/h/hospital_provider_spell_de.asp?shownav=1?query=%22spells%22&rank=14.28571&shownav=1.
313. NHS Digital. *NHS Business Definition: Surname*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/data_field_notes/p/pers/person_family_name_de.asp?shownav=1.
314. NHS Digital. *NHS Business Definition: Forename*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/data_field_notes/p/pers/person_given_name_de.asp?shownav=1.
315. Dusetzina, S.B., et al., *Linking data for health services research: a framework and instructional guide*. 2014: Agency for Healthcare Research and Quality (US), Rockville (MD).
316. Oxford Dictionaries. *Definition: Wild card*. 2016 [cited 2016 25/07/2016]; Available from: https://en.oxforddictionaries.com/definition/wild_card.
317. Gill, L. and M. Goldacre, *English national record linkage of hospital episode statistics and death registration records.*
318. NHS Digital. *NHS Business Definition: Date of birth*. 2016 [cited 2016 25/07/2016]; Available from:

http://www.datadictionary.nhs.uk/data_dictionary/data_field_notes/d/date_of_birth_(patient_identification)_de.asp?shownav=1.

319. NHS Digital. *NHS Business Rules Validation*. 2016 [cited 2016 24/12/2016]; Available from: http://www.datadictionary.nhs.uk/web_site_content/cds_supporting_information/commissioning_data_set_business_rules.asp?shownav=1?query=%2201%2F01%2F1900%22&rank=100&shownav=1.
320. NHS Digital. *NHS Business Definition: Sex*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/attributes/p/person/person_stated_gender_code_de.asp?shownav=1.
321. NHS Digital. *NHS Business Definition: Postcode*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/data_field_notes/p/pds/pds_postcode_(patient_address)_de.asp?shownav=1.
322. NHS England. *Data services for commissioners - assignment method flow chart (page 8)*. 2015 [cited 2016 17/12/2016]; Available from: https://www.england.nhs.uk/wp-content/uploads/2015/12/commissioner-assignment-process-flowchart-annotations.pdf.
323. Office for National Statistics. *NHS Data Field Notes: Postcode*. 2016 [cited 2016 17/12/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/data_field_notes/p/po/postcode_de.asp?shownav=1?query=%22no+fixed%22&rank=71.42857&shownav=1.
324. Office for National Statistics. *NHS Postcode Directory UK (August 2016)*. 2016 [cited 2016 17/12/2016]; Available from: https://ons.maps.arcgis.com/home/item.html?id=dc23a64fa2e34e1289901b27d91c335b.
325. Office for National Statistics. *2011 Lower Layer Super Output Areas*. 2013 [cited 2016 17/12/2016]; Available from: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/geography/products/census/lookup/other/index.html.
326. NHS Digital. *NHS Business Definition: Arrival Date*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/a/arrival_date_at_accident_and_emergency_department_de.asp?shownav=1.
327. NHS Digital. *NHS Business Definition: Arrival Time*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/a/arrival_time_at_accident_and_emergency_department_de.asp?shownav=1.
328. NHS Digital. *NHS Business Definition: Depart Date*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/a/accident_and_emergency_departure_date_de.asp?shownav=1.
329. NHS Digital. *NHS Business Definition: Depart Time*. 2016 [cited 2016 25/07/2016]; Available from: http://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/a/accident_and_emergency_departure_time_de.asp?shownav=1.
330. W3 Schools. *SQL Server Convert Function*. 2016 [cited 2016 30/12/2016]; Available from: http://www.w3schools.com/sql/func_convert.asp.
331. Farr Institute London. *Record Linkage Course*. 2013 [cited 2015 15/02/2015]; Available from: http://www.ucl.ac.uk/farr-short-courses/scfarr12.
332. NHS. *The NHS Strategic Tracing Service (NSTS)*. 2004 [cited 2016 30/12/2016]; Available from: http://webarchive.nationalarchives.gov.uk/20130502102046/http://connectingforhealth.nhs.uk/systemsandservices/nsts/docs/trace_out.pdf.

333. US Census. *Frequently Occurring Surnames from the 2010 Census*. 2017 [cited 2017 09/03/2017]; Available from: https://www.census.gov/topics/population/genealogy/data/2010_surnames.html.
334. National Records of Scotland. *100 Most common names National records of Scotland*. 2014 [cited 2017 09/03/2017]; Available from: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/births/popular-names/archive/100-most-common-surnames.
335. Evans, D., P. Gruba, and J. Zobel, *How to write a better thesis*. 2014.
336. SimMetrics. *String Metrics for SQL Server*. 2013 [cited 2015 03/05/2015]; Available from: https://sourceforge.net/projects/simmetrics/.
337. NHS England. *Process mapping, analysis and redesign*. 2017 [cited 2018 28/01/2018]; Available from: https://www.england.nhs.uk/improvement-hub/publication/improvement-leaders-guide-process-mapping-analysis-and-redesign-general-improvement-skills/.
338. Lee, J.T., et al., *The Effects of Pay for Performance on Disparities in Stroke, Hypertension, and Coronary Heart Disease Management: Interrupted Time Series Study.* PLoS ONE, 2011. **6**(12): p. e27236.
339. Pape, U.J., et al., *Disentangling secular trends and policy impacts in health studies: use of interrupted time series analysis.* Journal of the Royal Society of Medicine, 2013. **106**(4): p. 124-129.
340. Alshamsan, R., et al., *Effect of a UK pay-for-performance program on ethnic disparities in diabetes outcomes: interrupted time series analysis.* Ann Fam Med, 2012. **10**(3): p. 228-34.
341. Shiki, N., et al., *Unified Modeling Language (UML) for hospital-based cancer registration processes.* Asian Pac J Cancer Prev, 2008. **9**(4): p. 789-96.
342. HES Data Quality Team. *HES 2013-14 Annual Admitted Patient Care Data Quality Note*. 2015 [cited 2016 24/12/2016]; Available from: http://content.digital.nhs.uk/catalogue/PUB16725/nhs-mate-eng-2013-14-ip-qual-rep.pdf.
343. Weber, S.C., et al., *A simple heuristic for blindfolded record linkage.* Journal of the American Medical Informatics Association, 2012. **19**(e1): p. e157-e161.
344. Newman, D. *Episode of Care*. 2015 [cited 2019 05/12/2019]; Available from: https://healthcareitskills.com/episode-of-care/.
345. ICHNT. *Hammersmith A&E Closure*. 2014 [cited 2018 08/03/2018]; Available from: https://www.imperial.nhs.uk/about-us/news/response-to-hammersmith-and-fulham-councils-allegations-about-the-trust.
346. Bolisani, E.B.C., *Emergent Knowledge Strategies: strategic thinking in knowledge management*. 2019, [Place of publication not identified]: SPRINGER.
347. Campbell, E.M., et al., *Types of unintended consequences related to computerized provider order entry.* Journal of the American Medical Informatics Association : JAMIA, 2006. **13**(5): p. 547-556.
348. Warren, L.R., et al., *Improving data sharing between acute hospitals in England: an overview of health record system distribution and retrospective observational analysis of inter-hospital transitions of care.* BMJ Open, 2019. **9**(12): p. e031637.
349. Christen, P. and K. Goiser, *Quality and complexity measures for data linkage and deduplication*, in *Quality measures in data mining*. 2007, Springer. p. 127-151.