Mobile applications for the collection of patient self-reported data:

The case of depression during pregnancy

Jose Salvador **Marcano-Belisario**

Imperial College London
Department of Primary Care and Public Health
School of Public Health

A thesis submitted for the degree of Doctor of Philosophy
09 April 2019

Declaration of originality

This thesis presents my own work. Where applicable, the work of others is clearly 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. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Background

Mobile apps can operate as a delivery mode for many of the self-reported measures commonly used to gather patient data in healthcare. However, the incorporation of this technology in existing care pathways has the potential to influence the quality of the responses, as well as the routine activities of clinical services. This project evaluated the use of mobile apps for the delivery of depression screening scales and momentary questions during antenatal care.

Methods

A systematic review assessed the impact that apps as a delivery mode can have on the quality of self-report survey responses, compared to other delivery modes. A subsequent study evaluated the feasibility of using tablet computers for administering the Whooley questions and the Edinburgh Postnatal Depression Scale (EPDS) in the waiting area of antenatal clinics. Lastly, another study assessed the feasibility of using a bespoke app running on patients' own devices for the screening and monitoring of mood and depression in the last two pregnancy trimesters.

Results

The responses of surveys collected via apps are equivalent to those collected through traditional delivery modes. Moreover, it is feasible to use apps for depression screening in the waiting area of antenatal clinics. However, the adaptation of screening scales for use with an app could influence data quality, as suggested by a significant relationship between survey layout (i.e., the number of questions presented on a single screen) and scoring intervals on the EPDS. In addition, apps can assist in generating evidence of the diagnostic performance of screening scales in specific populations. Lastly, the development and deployment of apps for depression screening on patients' own devices must be sensitive to the needs of both patients and clinicians in order to minimise the likelihood of any potential detrimental effect.

Conclusions

Mobile technology is an innovative tool for supporting and complementing the role of healthcare practitioners, empowering patients, and enabling the reconfiguration of health services to better meet the needs of their users.

Table of Contents

Declaration of originality **2**

Copyright declaration **2**

Abstract **3**

Acknowledgments **9**

List of abbreviations **12**

List of Tables **15**

List of Figures **17**

1. Introduction **18**

1.1. Overview **18**

1.2. Clinical data and data quality **19**

1.2.1. Quantitative survey method 21

1.3. Mobile health **23**

1.3.1. Technical capabilities and portability 24

1.3.2. Operating System (OS) framework 25

1.3.3. Uptake and usage patterns 25

1.3.4. Methodological concerns 27

1.3.5. Attitudinal concerns 28

1.3.6. Practical concerns 29

1.4. Antenatal depression **29**

1.5. Thesis overview **32**

1.5.1. Aim 32

1.5.2. Chapter objectives 32

References (Chapter 1) **34**

2. Effect of apps as a delivery mode on the quality of self-administered survey responses: a systematic review **44**

2.1. Overview **44**

2.2. Rationale **46**

2.3. Background **46**

2.3.1. Apps and consumer mobile devices 49

2.4. Methods **50**

2.4.1. Aim 50

2.4.2. Inclusion criteria 51

2.4.3. Search methods 53

2.4.4. Data collection and analysis 54

2.5. Results **57**

2.5.1. Description of studies 57

2.5.2. Risk of bias 68

2.5.3. Effect of delivery mode 76

2.6. Discussion **87**

2.6.1. Limitations 92

2.6.2. Conclusions 93

References (Chapter 2) **95**

3. Tablet computer apps for depression screening in clinical settings **101**

3.1. Overview **101**

3.2. Rationale 102
3.3. Background 102
3.4. Methods 105
3.4.1. Aim 105
3.4.2. Study design 106
3.4.3. Sample size 106
3.4.4. Participants and participant recruitment 107
3.4.5. Technology and materials 110
3.4.6. Surveys 110
3.4.7. Survey layout 111
3.4.8. Study procedures 113
3.4.9. Statistical analyses 114
3.5. Results 115
3.5.1. Participants' responses to the Whooley questions and the EPDS 117
3.5.2. Completion time and requests for technical assistance 120
3.5.3. Smartphone and tablet computer ownership 121
3.6. Discussion 122
3.6.1. Limitations 127
3.6.2. Conclusions 128
References (Chapter 3) 129

4. Tablet computers for the assessment of the diagnostic performance of validated screening scales: the case of the Whooley questions 134
4.1. Overview 134
4.2. Rationale 135
4.3. Background 135
4.4. Methods 137
4.4.1. Aim 137
4.4.2. Participant recruitment and study procedures 137
4.4.3. Surveys 138
4.4.4. Technology and materials 139
4.4.5. Statistical analyses 139
4.5. Results 143
4.5.1. Characteristics of the participants 143
4.5.2. Participants' responses to the Whooley questions and the Edinburgh Postnatal Depression Scale 144
4.5.3. Diagnostic performance of the Whooley questions 145
4.6. Discussion 155
4.6.1. Limitations 158
4.6.2. Conclusion 158
References (Chapter 4) 159

5. Smartphones for the assessment and monitoring of antenatal depression in naturalistic settings using patients' own devices – Development of a study protocol 163
5.1. Overview 163
5.2. Rationale 165
5.3. Background 165
5.3.1. Retrospective versus momentary assessments: recall bias; sensitivity to temporal fluctuations; validity and reliability; and assessment reactivity 166
5.3.2. Patients' willingness to use their personal handsets: adherence and data completeness 167
5.3.3. Compatibility with existing services 168

5.4. Methods 169
5.4.1. Protocol of the feasibility study 169
5.4.2. Patient and public involvement strategy 181
5.5. Results 182
5.6. Discussion 182
5.6.1. Participant recruitment 182
5.6.2. Choice of self-reported scales 184
5.6.3. Momentary and retrospective assessments 184
5.6.4. Alerting system and duty of care protocols 185
5.6.5. Study outcomes 185
5.6.6. Notions of patient adherence 186
5.6.7. Conclusion 187
References (Chapter 5) 188

6. Smartphones for the assessment and monitoring of antenatal depression in naturalistic settings using patients' own devices – Study findings 192
6.1. Overview 192
6.2. Rationale 193
6.3. Background 193
6.3.1. Ecological momentary assessments as a strategy to promote participant adherence 194
6.3.2. Multiple definitions of participant adherence 194
6.4. Methods 195
6.4.1. Study design 196
6.4.2. Sample size 196
6.4.3. Study outcomes 197
6.4.4. Statistical analyses 198
6.5. Results 199
6.5.1. Baseline assessments 199
6.5.2. Adherence to sampling protocol 205
6.5.2.5. Drop-out rates 217
6.5.2.6. EPDS and EMA assessments 218
6.5.2.7. Alerts 219
6.5.2.8. Acceptance 221
6.6. Discussion 226
6.6.1. Participant adherence 227
6.6.2. Implementation challenges 231
6.6.3. Limitations 232
6.6.4. Conclusions 233
References (Chapter 6) 234

7. Discussion 238
7.1. Overview 238
7.2. Rationale 239
7.3. Discussion 239
7.3.1. Key contributions 239
7.4. Implications for clinical practice 248
7.4.1. Workload and workflows: the importance of external validity 248
7.4.2. Adaptation of a survey for use with an app: data quality 248
7.4.3. Integration of patient data into a consultation 249
7.4.4. Data processing and data visualisation 249
7.4.5. Locus of responsibility 251

7.4.6. Provision of basic infrastructure and information governance 251
7.4.7. Diversity and inequalities 251
7.5. Implications for research 253
7.5.1. Delivery mode effects 253
7.5.2. Importance of survey completion setting 253
7.5.3. Evidence-based implementation 254
7.5.4. Adherence 254
7.5.5. Pace of technology change 255
7.6. Limitations 255
7.7. Conclusions 257
References (Chapter 7) 258
List of articles published during doctoral studies 262
Appendix 1. Search strategy for use in Medline through OvidSP 263
Appendix 2. Search strategy for use in EMBASE through OvidSP 265
Appendix 3. Search strategy for use in PsycINFO through OvidSP 267
Appendix 4. Search strategy for use in IEEEXplore 269
Appendix 5. Search strategy for use in Web of Science 270
Appendix 6. Search strategy for use in CABI: CAB Abstracts 272
Appendix 7. Search strategy for use in Current Contents Connect 274
Appendix 8. Search strategy for use in ACM Digital 276
Appendix 9. Search strategy for use in ERIC 277
Appendix 10. Search strategy for use in Sociological Abstracts 278
Appendix 11. Search strategy Health Management Information Consortium 279
Appendix 12. Search strategy for use in CENTRAL 281
Appendix 13. Search strategy for use in ClinicalTrials.gov 283
Appendix 14. Search strategy for us in ProQuest Dissertation and Theses 284
Appendix 15. Initial Search Results 285
Appendix 16. Update search results 286
Appendix 17. Characteristics of Excluded Studies 287
Appendix 18. Characteristics of Ainsworth and colleagues (2013) 295
Appendix 19. Characteristics of Brunger and colleagues (2015) 297
Appendix 20. Characteristics of Bush and colleagues (2013) 299
Appendix 21. Characteristics of Garcia-Palacios and colleagues (2014) 301
Appendix 22. Khraishi and colleagues (2012) 303
Appendix 23. Characteristics of Kim and colleagues (2014) 305
Appendix 24. Characteristics of Lamber and colleagues (2012) 307
Appendix 25. Characteristics of Newell and colleagues (2015) 310

Appendix 26. Characteristics of Salaffi and colleagues (2013) 313
Appendix 27. Characteristics of Schemmann and colleagues (2013) 315
Appendix 28. Characteristics of Sigaud and colleagues (2014) 316
Appendix 29. Characteristics of Stomberg and colleagues (2012) 318
Appendix 30. Characteristics of Sun and colleagues (2013) (A) 321
Appendix 31. Characteristics of Sun and colleagues (2013) (B) 322
Appendix 32. Socio-demographic survey 323
Appendix 33. Whooley questions 325
Appendix 34. Edinburgh Postnatal Depression Scale 326
Appendix 35. Momentary questions 328
Appendix 36. Contextual questions 329
Appendix 37. Acceptance survey 330
Appendix 38. Overview of findings with references 331
Survey completion setting 331
Mediators of delivery mode effects 332
Data collection in clinical settings 333
Data collection in naturalistic settings 334
References 336

Acknowledgments

This work was funded by the Global eHealth Unit through the NIHR Imperial BRC. Thank you to Azeem Majeed for welcoming me into the Department of Primary Care and Public Health. Thank you to my supervisors Dr Josip Car and Dr John O'Donoghue.

Thank you to Laura Gunn for her mentorship.

Thank you to Christine Franey for listening to me on so many occasions.

Thank you to Aleš Porčnik, Andreja Saje and Jam Jamšek for their help with my systematic review.

Thank you to Antoinette McNulty at the CRN for Northwest London for guiding me through the governance and logistic requirements of conducting research in the NHS, and for her sound advice.

Thank you to Nikolaos Mastellos, Martin Cohn, Yvonne Tul, Laura Pearse, Geraldine Landers, Joanna Girling, Marie O'Connell, Cathie Melvin, Tara Pauley, Charlotte Clayton, Anna Fleming, Jane Radford, Anne Hogg, Angela Loughlin, Mark Johnson, Roshni Patel, Carmella Martella, Minna Meritahti, Claire Fairhurst, Tracey Taylor, Rachel Newport, Zainab Sarwar Yasin, Alison Loftus, Sally Dunn, Siobhan Limerick, Helen Riding, Justine Norman, Helen Ginelly, Mary Day, Gemma Loebenberg, Sophie Hannay, Saloni Dosani, Shilpi Mehra and Tara Harvey for supporting and facilitating my work.

Thank you to Susan Ayers and Alex Thornton at City University for sharing their knowledge and expertise.

Thank you to Ajay Gupta and Paul Ramchandani for their advice, time and feedback.

Thank you to Gavin Doherty and Kevin Doherty at Trinity College London for their collaborative spirit.

Thank you to Kit Huckvale and Mark Larsen for their mentorship, stimulating conversations, and more.

Thank you to Paul Miller and Mary Shaw for their support and advice during difficult times.

A special thank you to my supervisor Cecily Morrison for her continuous support, guidance and encouragement, and for being responsible for some of the most enjoyable aspects of this PhD.

To Nahir and Eyra for helping me to get here.

To Miriam, for being by my side throughout this journey.

List of abbreviations

3G:	Third generation mobile telecommunications
4G:	Fourth generation mobile telecommunications
ADL:	Activities of daily living
Apps:	Mobile applications
AUC:	Area Under the Curve
BASDAI:	Bath Ankylosing Spondylitis Disease Activity Index
BASFI:	Bath Ankylosing Spondylitis Functional Index
CAS:	Color Analog Scale
CES-D:	Center for Epidemiologic Studies Depression Scale (CES-D)
CI:	Confidence intervals
CLAHRC:	Collaboration for Leadership in Applied Health Research and Care
CSO:	Clinical studies officer
DAR5:	Dimensions of Anger 5
DRRI-US:	Deployment Risk and Resilience Inventory-Unit Support
EMA:	Ecological momentary assessment
EORTC QLQ – 30:	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire C-30
EPDS:	Edinburgh Postnatal Depression Scale
FOR:	False Omission Rate
FPS-R:	Faces Pain Scale Revisited
GP:	General practitioner
GPRS:	General packet radio service
GPS:	Global positioning system
GRADE:	Grading Recommendations Assessment, Development and Evaluation
HAQ:	Health Assessment Questionnaire
HIV:	Human immunodeficiency virus
ICC:	Intra-class correlation
iHOT-12:	Short International Hip Outcome Tool

iOS:	iPhone Operating System
IPSS:	International Prostate Symptom Score
ISPOR ePRO:	The International Society for Pharmacoeconomics and Outcomes Research electronic Patient-Reported Outcome
MD:	Mean difference
mHealth:	Mobile health
MID:	Minimally important difference
MMS:	Multimedia messaging service
NHS:	National Health Service
NICE:	National Institute for Health and Care Excellence
NPV:	Negative Predictive Value
NRS:	Numerical rating scale
OR:	Odds ratio
OS:	Operating system(s)
PANSS:	Positive and Negative Syndrome Scale
PDA:	Personal digital assistant
PHQ-9:	Patient Health Questionnaire
PIL:	Participant information leaflet
PPI:	Patient and Public Involvement
PPV:	Positive Predictive Value
PROM:	Patient-reported outcome measure
PTSD-Checklist:	Post-Traumatic Stress Disorder Checklist
RCT:	Randomised controlled trial
RFQ:	Regulatory Focus Questionnaire
ROC:	Receiver Operating Characteristics
R-SIS:	Revised Suicidal Ideation Scale
SD:	Standard deviation
SEM:	Standard error of the mean
SMS:	Short messaging service
UK:	United Kingdom
US:	United States
USE:	Usefulness, Satisfaction and Ease of Use Questionnaire

VAS:	Visual analogue scale
WHO:	The World Health Organization

List of Tables

Table 1. Advantages and disadvantages of interview-administered and self-administered surveys (adapted from Carter Y, Shaw S, Thomas C, p. 13) 48
Table 2. Participant inclusion and exclusion criteria 108
Table 3. Participants' socio-demographic characteristics 116
Table 4. Number of participants according to their answers to the Whooley questions 117
Table 5. Participants' overall mean and median EPDS scores, and by survey layout 119
Table 6. Overall number of participants at each EPDS scoring interval, and by survey layout 119
Table 7. Number of participants according to their answers to question 10 of the EPDS, and by survey layout 120
Table 8. Overall number of participants requesting technical assistance, and by survey layout 121
Table 9. Number and type of technical assistance requested by participants 121
Table 10. Participants' socio-demographic characteristics 143
Table 11. Number of participants at each EPDS scoring interval, and by pregnancy trimester, obstetric history, ethnic background and previous history of depression 145
Table 12. Number of participants according to their answers to the Whooley questions and EPDS scores (EPDS threshold: 13 points) 146
Table 13. Number of participants according to their answers to the Whooley questions and EPDS scores (EPDS threshold: 15 points) 146
Table 14. Number of participants according to their answers to question one of the Whooley questions and EPDS scores (EPDS threshold: 13 points) 147
Table 15. Number of participants according to their answers to question one of the Whooley questions and EPDS scores (EPDS threshold: 15 points) 147
Table 16. Number of participants according to their answers to question two of the Whooley questions and EPDS scores (EPDS threshold: 13 points) 147
Table 17. Number of participants according to their answers to question two of the Whooley questions and EPDS scores (EPDS threshold: 15 points) 147
Table 18. Number of participants who answered Yes to both Whooley questions according to their EPDS scores (EPDS threshold: 13 points) 148
Table 19. Number of participants who answered Yes to both Whooley questions according to their EPDS scores (EPDS threshold: 15 points) 148
Table 20. Summary of the diagnostic performance of the Whooley questions in relation to two EPDS scoring thresholds (i.e., 13 and 15 points) 153
Table 21. AUC of the Whooley Questions for different EPDS thresholds 153
Table 22. Participant inclusion and exclusion criteria 171
Table 23. Alerting system 179
Table 24. Participants' socio-demographic characteristics 200
Table 25. Proportion of participants who did and did not download, install and activate BrightSelf by group allocation 202
Table 26. Proportion of participants who did not activate BrightSelf by group allocation and locality 203
Table 27. Overall proportion of operating systems, and by group allocation 203

Table 28. Proportion of respondents by group allocation, amongst those who downloaded, installed and activated BrightSelf 204
Table 29. Proportion of respondents by group allocation, amongst those randomised into the study 204
Table 30. Number of participants who completed all the expected assessments during each sampling period 206
Table 31.Number of participants who completed all the expected assessments during each sampling period, after having completed the assessments in the immediately preceding sampling periods 208
Table 32. Adherence rate by allocation group, defined as the number of participants who completed all the expected assessments as a proportion of the number of participants randomised into the study 209
Table 33. Number of participants who completed at least one expected assessment during each sampling period, regardless of whether they completed the preceding sampling periods 210
Table 34. Number of participants who completed at least one expected assessment in each sampling period, dependent on having completed the preceding sample periods 212
Table 35. Adherence rate by group allocation for the whole duration of the study, defined as the number of participants who completed at least one of the expected assessments in each sampling period for all sampling periods as a proportion of the number of participants randomised into the study 213
Table 36. Adherence rate by group allocation, defined as the number of participants who completed at least one expected assessment as a proportion of the number of participants randomised into the study 214
Table 37. Number of participants according to the number of assessment periods they completed (EPDS assessments only) 215
Table 38. Number of participants who completed one assessment period according to the assessment period they completed (EPDS assessments only) 215
Table 39. Number of participants who completed two assessment periods according to the assessment periods they completed (EPDS assessments only) 216
Table 40. Number of participants who completed three assessment periods according to the assessment periods they completed (EPDS assessments only) 216
Table 41. Number of participants who completed four assessment periods according to the assessment periods they completed (EPDS assessments only) 217
Table 42. Number of participants who completed five assessment periods according to the assessment periods they completed (EPDS assessments only) 217
Table 43. Drop-out rates by group allocation 218
Table 44.Number of alerts by type and group allocation 220
Table 45. Number of participants in each experimental group according to the number of alerts the triggered 220

List of Figures

Figure 1. PRISMA flow diagram 58
Figure 2. Risk of bias graph for RCTs 69
Figure 3. Risk of bias summary for RCTs 69
Figure 4. Risk of bias graph for crossover trials 72
Figure 5. Risk of bias summary for crossover trials 73
Figure 6. Study flow 109
Figure 7. Scrolling layout through SNAP® surveys (screenshot) 112
Figure 8. Paging layout through SNAP® surveys (screenshot) 113
Figure 9. Distribution of EPDS scores in the sample of participants 118
Figure 10. ROC curves of the Whooley Questions for the following EPDS thresholds: 12, 13 and 14 points 154
Figure 11. BrightSelf - Check Back (Screenshot) 174
Figure 12. BrightSelf - Check In (Screenshot) 175
Figure 13. Sampling protocols according to type of assessment strategy 178
Figure 14. Number of participants who completed all the expected assessments during each sampling period. 207
Figure 15. Number of participants who completed at least one expected assessment during each sampling period 211
Figure 16. Number of participants agreeing/disagreeing with the statement: the app was easy to use 221
Figure 17. Number of participants agreeing/disagreeing with the statement: I found using this app an engaging experience 222
Figure 18. Number of participants agreeing/disagreeing with the statement: I found the assessments useful 223
Figure 19. Number of participants agreeing/disagreeing with the statement: I was asked to provide reports… (related to their frequency) 224
Figure 20. Number of participants agreeing/disagreeing with the statement: the experience of using this app met my needs 225
Figure 21. Number of participants in the Retrospective plus momentary group agreeing/disagreeing with the statement: I found it useful to compare reports made right now and those made over the past 7 days 226
Figure 22. Conceptual framework of the effect that technology for the collection of clinical data has on data quality 246

1. Introduction

1.1. Overview

Data play a key role in healthcare. The success of activities such as clinical-decision making, population health assessments, planning and commissioning of services, and health policy rely heavily on the quality of their source data. Data quality in turn, is largely influenced by the choice of data collection method. The quantitative survey method is one of the most commonly used methods in healthcare. Through the administration of surveys, it provides the means to produce quantified measures of patients' symptoms, behaviours, attitudes, feelings, opinions and knowledge. However, the quality of survey responses is determined by the complex interaction between surveys, patients and the delivery mode. This interaction is also influenced by the situations in which survey completion takes place. Apps running on smartphones and tablet computers offer a new compelling delivery mode for health-related surveys. As such, they create new conditions for the interaction between surveys, patients and delivery mode that must be assessed. Antenatal depression represents an excellent opportunity to assess these conditions given its public health burden, the highly-structured nature of antenatal care in England and other developed countries, and the existence of validated surveys for its recognition.

This introductory chapter explores notions of the quality of clinical data in relation to the quantitative survey method. It then outlines the features that enable apps running on smartphones and tablet computers to operate as a delivery mode for health-related surveys. Subsequently, it presents the rationale for choosing antenatal depression to assess the potential effect of apps on survey responses. This chapter concludes with an outline of the thesis.

1.2. Clinical data and data quality

Clinical data are generated mainly during care encounters between patients and their providers (typically captured in health records), as well as during research activities or large-scale surveys [1, 2]. The primary purpose of these data is to inform patients' direct care [3]. This usually involves diagnostic and therapeutic decisions, monitoring disease progression and treatment response, and the identification of behavioural risk factors. Individual-level data can also be aggregated into larger data sets (e.g., national data sets or disease registries) for secondary purposes [1, 3, 4]. These include assessing the health status of populations; planning and commissioning services; evaluating the effectiveness, efficiency, safety and acceptability of services and interventions; guiding the implementation and evaluation of public health programmes; and informing health policy [2, 3, 5]. Since the outcomes of these primary and secondary activities affect the health of individuals and populations, it is important to ensure the quality of their source data. If the source data are inaccurate or incomplete, decisions and recommendations will be biased, misleading and potentially harmful [6-10].

Ensuring the quality of clinical data can be challenging. Although data quality has been recognised as a multi-dimensional construct, there is no universal agreement regarding its core defining dimensions [11, 12]. Neither is there a widely accepted standardised framework for conducting comprehensive data quality assessments (including benchmarking) both within and between organisations [11]. Consequently, ad hoc measures are often implemented. These measures tend to capture stakeholders' subjective perception of what data quality is supposed to mean, as well as descriptive metrics related to the state of a dataset and the task at hand [12]. Despite this variability, metrics such as completeness, accuracy and timeliness are commonly used as indicators of data quality in healthcare, as well as concepts such as validity and reliability. When conceptualised as such, it becomes necessary to understand the factors that can influence and modify these data quality indicators.

Data creation (and more specifically data collection) is widely acknowledged as one of the key determinants of data quality [6, 13, 14]. To better illustrate how decisions made during this initial stage can influence data quality it is helpful to use a data life cycle framework. Many

frameworks have been proposed [15], which differ in the number of stages they consider. However, they invariably contain a data creation stage, with data collection as one of its key steps. Here, the framework proposed by the UK Data Service will be used [16].

The UK Data Service's Research Data Lifecycle framework identifies 6 stages: (i) creating data; (ii) processing data; (iii) analysing data; (iv) preserving data; (v) giving access to data; (vi) and re-using data [16]. To create data, relevant stakeholders (e.g., researchers or clinicians) must first decide on the objective(s) they wish to achieve (e.g., answering a research question or implementing a screening programme). Thereafter, this stage can be divided into a planning phase, and an implementation phase. As part of the former, stakeholders will identify the requirements needed to accomplish their objective(s): target population; type of data required; data collection protocol (in terms of frequency and duration); data management plans; data storage systems; plans for consent and for sharing; and regulatory frameworks amongst other requirements [6, 8, 15-18]. Decisions on these requirements will then inform the choice of data collection method, from a wide range of techniques such as direct observation, quantitative surveys, qualitative interviews, ethnography, and registration [19]. Data collection is thus the implementation of a chosen method in order to gather data that are (i) relevant to the objective(s) and the proposed plan; (ii) accurate; (iii) complete, (iv) standardised across sites and across similar projects, (v) efficient for data recording and processing, and (vi) appropriate for analysis and decision-making [6, 13].

The choice of data collection method is then one of the most important decisions in the data lifecycle, as it will determine the suitability of the data for achieving the planned objective(s) [13]. Moreover, the practical implementation of a data collection method can lead to errors that will affect data quality [6, 13]. Since errors made at this point are difficult to detect and correct at future stages, quality assurance should be an integral part of data collection [6]. However, the factors that can lead to errors in the data vary according to the data collection method. For this reason, the choice of quality assurance measures must be specific to a particular method or combination of methods.

The focus of this work will be on issues of data quality related to the quantitative survey method.

1.2.1. Quantitative survey method

The quantitative survey method is one of the most commonly used data collection methods in healthcare [20, 21]. It consists in administering surveys to a sample of respondents; these surveys contain questions or statements that assess respondents' symptoms, perceived health status, attitudes, beliefs, behaviours, feelings, knowledge and/or opinions [9, 21]. The key assumption is that responses given to these items can be converted into quantifiable, objective measures of the construct under evaluation [21]. These estimates can then be used to make statistical inferences about respondents and the population from which they were sampled [7, 20, 21].

The quantitative survey method is typically associated with large-scale population surveys (such as the Health Survey for England, or the Adult Psychiatric Morbidity Survey). However, it is also widely used for assessing symptoms or diagnosing patients (e.g., the Hospital Anxiety and Depression Scale), exploring lifestyle risk factors (e.g., Alcohol Self-Assessment Tool in National Health Service [NHS] clinics) or assessing quality of life (e.g., the Quality of Life Scale) amongst others.

Data quality in relation to the quantitative survey method must be explored in relation to survey responses and survey completion as a process. Regarding survey responses, the survey error model provides a useful framework [22, 23, 24]. Survey error can be of two types: errors of measurement and errors of representation. The former refers to discrepancies between the responses given to a survey, and the true value of the underlying attribute [23]. This type of error is linked to the concepts of validity and reliability (also referred to as psychometric properties), which form the basis of one of the main distinctions between surveys: validated surveys and non-validated surveys. Validated surveys are those that have had their psychometric properties tested and are thought to provide valid and reliable estimates of the constructs of interest. Non-validated surveys, on the other hand, are those that have not been subjected to a formal testing process. As a result of this, there is no guarantee that they can provide reliable estimates [25]. Errors of representation refer to discrepancies between the statistics estimated on a sample of individuals, and those estimated on the target population [22, 24]. Examples of this type of error include the appropriateness of sampling frames, and non-response bias.

Survey responses however, are the result of a complex interaction. Generating responses requires a series of cognitive processes such as question comprehension, memory retrieval, judgement and estimation, and reporting [22, 26]. Factors operating concurrently with these cognitive processes can then influence how responses are generated. These factors include the survey instrument, patients themselves and the delivery mode [7, 18, 21, 26-28]. For example, the wording of questions and an individual's literacy levels will determine the ease with which questions are comprehended. The type of response required (e.g., factual, estimation of a time interval, or an attitudinal judgement) will determine the cognitive processes required to generate the answer (e.g., recall or reconstruction), and the type of response biases that might arise [22]. For example, depending on the nature of the survey, some people may edit their initial responses to present themselves in a more favourable light (i.e., social desirability bias) [26]. Moreover, the channels (e.g., visual or auditory) through which questions and response options are presented will determine the load on an individual's cognitive resources [22, 26]. These examples illustrate the complexity of the interaction between these three factors, and the difficulty of isolating the effect of any one of them.

The influence exerted by the survey instrument, patients and the delivery mode depends on how the quantitative survey method is implemented. It is in this implementation that the other facet of data quality becomes apparent: survey completion as a process. In healthcare, the administration of surveys is typically linked to a specific objective (e.g., informing a therapeutic decision, or gather information about a patient's lifestyle). For this reason, it is important that survey completion and the intrinsic requirements of the intended objective(s) are aligned. For example, answering all the questions so that an overall score can be calculated, or providing multiple data points to illustrate a trend in clinical or behavioural parameters. In addition, survey completion ought to facilitate the activities of all the stakeholders involved in a clinical activity, including both patients and clinicians. For example, if the survey can only be actioned once an overall score has been calculated, then practitioners must have enough time between receiving the completed survey and a consultation to generate this score. This involves planning carefully how surveys will be circulated to and collected from patients.

In practice, surveys are typically implemented as part of a highly-structured process where patients (or respondents) represent repositories of clinical data. Through their interaction with these patients (usually in clinical settings), trained professionals can use surveys to extract data in a format that is suitable for future processing and consumption. Validated surveys are often used to standardise this process, as they typically yield a score that suggests a possible diagnosis and a pre-defined course of action. Alternatively, surveys can be given directly to patients for them to complete independently. This still requires trained professionals to provide some level of support (such as clarification of questions, or reassurance about the potential implications of the results), and to check and score patients' responses. Surveys have traditionally been delivered using paper, although electronic devices are increasingly becoming the norm.

These conditions provide a relatively controlled environment for surveys, patients and the delivery mode to interact. They also enable the implementation of quality assurance measures with relative success. These measures include the use of standardised methods and protocols, piloting data collection instruments, taking multiple observations, use of validation rules for data entry, and controlled vocabulary or code lists to minimise manual data entry amongst others [13]. However, recent developments in mobile technology can affect this interaction not only by modifying one of its key factors (i.e., delivery mode), but also by changing the conditions under which survey completion can take place.

1.3. Mobile health

The use of mobile technology for medical and public health practice is referred to as mobile health (mHealth) [29, 30]. A key feature of mHealth is taking advantage of the core features of devices such as mobile phones (both feature phones and smartphones), personal digital assistants (PDA), tablet computers, and (wearable) sensors or monitoring devices with wireless capabilities [29, 31]. For general-purpose devices, these core features include voice communication, short messaging service (SMS), multimedia messaging service (MMS), general packet radio service (GPRS), third and fourth generation mobile telecommunications (3G and 4G, respectively), global positioning system (GPS), wireless technology (including Bluetooth), and built-in sensors (such as gyroscopes and accelerometers) [29, 31]. The core features of specialised monitoring devices consist of a combination of wireless

telecommunication capabilities with biochemical, physiological or biometric sensors (e.g., wireless blood glucose monitors) [31]. Overall, these devices can enable initiatives such as improved data collection at point of care, delivery of care and health promotion activities, patient education, health monitoring and surveillance, improved communication (between patients and health services, between healthcare professionals, and between services or departments), and improved access to information for healthcare professionals at point of care [29, 30, 32].

Since mHealth encompasses a wide range of (i) devices and (ii) interventions, it is important to be specific about these aspects. Here, the term mHealth refers to the use of mobile applications (apps) running on consumer general purpose devices (i.e., smartphones and tablet computers) for the collection of clinical data through survey instruments. To better understand how these devices can transform the traditional implementation of the quantitative survey method, and the impact they can have on survey responses, it is important to explain the characteristics that enable them to operate as data collection tools: (i) technical capabilities and portability, (ii) operating system (OS) framework, and (iii) high uptake and unique usage patterns.

1.3.1. Technical capabilities and portability

Concerning their technical capabilities, smartphones and tablet computers are equipped with screens that can display information in several formats: text, images, animations and videos, and audio. Similarly, various input methods (i.e., touchscreen gestures, on-screen keyboards, cameras, microphones, Bluetooth connectivity with external devices, and built-in sensors) support the capture of data in multiple formats. This could enrich the data collection process by enabling the collection of the specific data items of interest (e.g., survey responses) and complementing them with contextual information such as time of day or location [33]. Given their processing power, smartphones and tablet computers can deal with complex data processing requirements; they can also deal with relatively large volumes of data as a result of their built-in storage capacity (which in some cases can be expanded) [34]. In addition, the networking capabilities of these devices allow them to transfer data without temporal or geographical constrains [10, 18, 34-36]. Not least, their dimensions make them easily transportable.

1.3.2. Operating System (OS) framework

One the central features of the new OS framework for smartphones and tablet computers is its focus on apps. Through the OS, apps can access the technical capabilities of a general-purpose device and enable it to perform specialist functions. In addition, this new OS framework has changed traditional software distribution models. Apps can be installed on a smartphone or tablet computer by the device manufacturer. However, they can also be developed by third parties and made available through marketplaces such as the AppStore [37] or Google Play [38]. Users can easily download these apps from these marketplaces, and directly install them on their devices [8, 35, 36]. Future changes to the content and functionality of an app can be deployed as updates through the same marketplaces. This distribution model offers advantages in terms of flexibility and convenience.

1.3.3. Uptake and usage patterns

Smartphones and tablet computers have been widely accepted worldwide. In 2017 in the United Kingdom (UK) alone, approximately 76% of adults reported owning a smartphone, and approximately 58% reported owning a tablet computer [39, 40]. In addition, smartphones have become the primary device used by 66% of adults in the UK to access the internet [39]. These devices (particularly smartphones) have also resulted in unique usage patterns in terms of the frequency and duration, location and types of interactions. Smartphone usage is typically characterised by short bursts (of around 15 seconds) of habitual interactions [41, 42]. For approximately 69% of adult users in the UK, these interactions start within 30 minutes of waking up, and for another 58% end within 30 minutes of going to sleep [40]. People have reported using their devices always or very often in a wide range of situations: at work, socialising with family or friends, walking, shopping, commuting and whilst watching a film [40, 41]. Furthermore, apps have made smartphones and tablet computers extremely versatile. Therefore, a single device can be used for multiple purposes: accessing the internet, controlling one's home, keeping diaries or calendars, shopping, games and other forms of entertainment, navigating the environment, taking selfies, accessing financial services, and social media [39, 40, 43]. For some people smartphones and tablet computers have become an essential part of their lives [32, 43-45].

However, there are differences in ownership and usage patterns between segments of the population. Smartphone ownership tends to decline with age, level of deprivation, or with disability (including visual, mobility and hearing impairments, as well as learning disabilities) [46]. Regardless of ownership, individuals from hard-to-reach groups (such as ex-offenders or transient occupants) might be less willing to use smartphones in certain situations due to concerns about privacy, for example [47]. In the context of mHealth, not acknowledging these differences could widen existing health inequalities if these personal and socio-demographic characteristics are also associated with increased health risks.

The combination of these characteristics enables smartphones and tablet computers to operate as tools for the administration of surveys in a way that is convenient for both clinicians (or researchers) and patients. Apps offer the opportunity to administer surveys to collect a wide range of behavioural and symptom data in traditional formats (i.e., text). In the case of validated surveys, the processing power of consumer devices allows them to deal with complex scoring requirements. Moreover, the portability, connectivity and usage patterns of smartphones and tablet computers can support the collection of momentary, experiential and ecologically valid data [48], and offer the possibility of responding to the data collected in real time. In terms of scalability, a resource-intensive period of initial app development could be followed by a relatively low-cost, large-scale distribution to a diverse and (potentially) geographically dispersed user base. Users' familiarity with apps and their devices could reduce training requirements. For these reasons, apps running on consumer devices are already being used in healthcare. For example, for monitoring signs and symptoms of asthma [49], tracking physical activity [50], and for depression screening and other mental health conditions [51, 52].

Nonetheless, the rapid adoption of apps as survey administration tools does not immediately guarantee the quality of the resulting responses. Although quality assurance measures common to other electronic devices can be implemented with apps, the characteristics of this technology can influence survey responses in novel ways. This has contributed to raise some methodological, attitudinal and practical concerns.

1.3.4. Methodological concerns

From a methodological perspective, the main concerns relate to the validity and the reliability of surveys, response biases, patient engagement, and standardisation of clinical data.

Apps support the administration of validated surveys. However, the conditions under which these measures can be completed, including the frequency with which they can be administered, may differ from the conditions under which the original instrument was validated. Therefore, it might be necessary to assess whether the responses collected through validated instruments delivered via apps are equivalent to those collected using traditional channels. These assessments should also consider the potential influence of the data completion setting (e.g., clinical, community, at-home, commuting, or at work) on the properties of the survey. Furthermore, apps can enable the collection of unstructured, non-validated data in multiple formats. These data need to undergo a thorough validation process before they can be used to inform clinical decisions.

The interaction between the survey instrument, patients and the delivery mode can influence, and sometimes bias, patients' responses. Phenomena such as recall bias, social desirability, probability judgements, anchoring bias, and contextual effects have repeatedly been reported in the literature [26, 48, 53-56]. However, since apps running on consumer devices is a relative new development, it is still unclear how they could affect the size and direction of these phenomena.

Moreover, apps could be used as the patient-facing interface of a clinical system, through which patients can enter data directly into the system. For this reason, patient engagement has become a prerequisite for the effectiveness of these systems [57, 58]. Although numerous definitions have been proposed over the years, *effective engagement* provides a useful way of conceptualising this construct [58]. Defined as such, engagement refers to users' experience of the technology in relation to its purpose, and the social and clinical context in which it is deployed [58]. The key advantage of this definition is that it offers a wide range of methods for measuring engagement, and for promoting it through a number of design choices. Well-designed systems can attract patients' attention and maintain their long-term

engagement, minimise the disruption of the data collection process to their daily routines, and reduce the potential negative influence of response biases.

Lastly, one of the main ways in which delivery modes can influence survey responses is through the interaction between a patient and the input method. The technical specifications of a device can affect this interaction, and thus data entry. For example, screen size can affect patients' judgements on a sliding scale, or can result in frustration when attempting to enter responses through checkboxes, radio buttons or an on-screen keyboard. This can make the standardisation of survey responses across patients difficult given the wide range of devices (with differing technical specifications) that are available on the market. Therefore, efforts are needed to determine if this could be a genuine source of survey error.

1.3.5. Attitudinal concerns

An additional concern is the attitude of both patients and clinicians towards the use of apps for the administration of surveys. In this respect, a key factor is patients' willingness to run what is effectively a clinical tool on their personal devices [59]. This factor could greatly influence patient compliance with surveys, and result in response biases or incomplete data.

Moreover, these technologies are sometimes portrayed in the media as a key step toward the automation of data collection and of clinical decision-making. In addition, they are often developed without considering the potential impact on healthcare workflows [58]. Therefore, these technologies could threaten the sense of professional identity and disrupt the activities of healthcare professionals. From a data quality perspective, this could affect clinicians' endorsement of the technology and its successful incorporation within existing care pathways, as well as influence patients' perception. This could ultimately lead to response biases or poor patient adherence. For this reason, efforts are needed to ensure that mobile systems meet the expectations of their intended users, and help them to achieve their goals [58].

1.3.6. Practical concerns

One final aspect is the practical implications of introducing apps into existing care pathways, particularly given the pace of technological change [60]. Apps can enable situations in which patients cannot access immediate support from a trained professional. For example, to clarify questions, raise concerns about privacy and confidentiality, or to confirm the purpose of a survey. This could result in response biases, thus raising questions about the validity and reliability of the responses collected through them. Moreover, some patients might use their devices in limited ways. If assumptions based on traditional usage patterns are used to inform data collection protocols, this could result in increased disruption to patients' routines and affect patient engagement. Lastly, technology changes at an incredibly fast pace. New devices, OS and software development standards are released on a yearly basis. This means that relevant stakeholders must ensure that their apps are compatible with the latest developments in order to ensure an optimal user experience, and the privacy and security of the data.

Overall, the potential effect of smartphone or tablet computer apps on survey responses offer new research opportunities. These can be explored in relation to specific target populations and specific disease areas.

1.4. Antenatal depression

Mobile technology has been used for collecting information about signs and symptoms, medication taking and lifestyle risk factors across various disease areas, as well as diverse settings [49, 50, 61-66]. More recently, the role of apps for the assessment and screening of mental health conditions has also been recognised [51, 59, 67-70].

In the UK, mental health problems are the main contributor to the burden of disease (approximately 28%), affecting up to one in four people in any given year [71]. These disorders are also key drivers of disability and are a considerable source of economic burden (between £70 and £100 billion a year) [71]. In addition, there are strong links between mental health and physical health, which can significantly affect patient recovery [72].

Smartphone and tablet computer apps might prove particularly useful for identifying and monitoring these conditions in non-mental health settings (e.g., primary care, and antenatal care), as case-finding and disease progression monitoring often rely on validated surveys. This approach however, should be followed with caution as issues related to the use of validated surveys in mental health are still relevant regardless of delivery mode. Although ideal for standardising clinical assessments (and for providing non-specialist practitioners with clear actionable recommendations), validated surveys leave little room for clinical judgement [73]. In these cases, the observed prevalence has been found to be a product of the underlying distribution of symptom levels in the target population, rather than the properties of the survey [73]. For example, Levis and colleagues found that compared with semi-structured interviews (which allow for clinical judgement), participants with low levels of depression symptoms were more likely to be identified as having major depression when assessed through a validated survey and vice versa [73]. Therefore, although apps can facilitate the administration and management of these surveys, they should not replace clinical judgement.

In addition, apps can support the implementation of methods, such as ecological momentary assessment (EMA), that are better suited to the study of mental health phenomena and their fluctuations over time [74, 75].

The term mental health disorders encompasses a wide range of conditions that differ in their clinical characteristics, course and prognosis. This project will focus on depression that occurs during pregnancy (or antenatal depression). One reason for this choice is its public health burden [71, 76-80]. The point prevalence of antenatal depression varies between 7% and 12% of pregnant women, depending on factors such as pregnancy trimester, obstetric history (including current obstetric complications and previous miscarriages) and history of mental health disorders [79, 81-83]. In addition, period prevalence estimates suggest that as many as 12.7% of women will experience at least one episode of major depression at any point during their pregnancy [79, 82]. Untreated antenatal depression can increase the likelihood of engaging in unhealthy practices (i.e., unhealthy diet, substance abuse, failure to enrol in antenatal care), and the risk of self-harm and postnatal depression (including more severe forms of this condition) [76-78, 84]. In addition, antenatal depression can affect foetal development and result in adverse birth outcomes (such as preterm birth and low birth

weight), and has been identified as an independent risk factor for a child's behavioural, cognitive and emotional development [76-78, 85-89]. Moreover, antenatal depression can also affect a woman's relationship with her partner and relatives [90]. Research indicates that there is no difference between pregnant and non-pregnant women in the prevalence or incidence of depression [91]. However, the rate of diagnosis and provision of treatment appears to be lower in pregnant women [91]. Up to three-quarters of pregnant women meeting diagnostic criteria for depression (and anxiety) are not identified, and only one in ten of those needing further treatment accesses it (despite the availability of adequate help in primary care and other healthcare settings) [92].

The short- and long-term effects of antenatal depression, coupled with the direct and indirect financial costs to the healthcare system, account for a large proportion of the societal costs of perinatal mental health disorders [93]. In the UK, these costs have been estimated at £8.1 billion for each one-year cohort of birth, with the average cost of each case of perinatal depression (both antenatal and postnatal) estimated at £74k [90]. Approximately 72% of these costs can be attributed to the adverse effects experienced by the child [90].

Another reason for focusing on antenatal depression, is that antenatal care tends to have clearly defined pathways during which women experience repeated encounters with their healthcare providers. In England, the National Institute for Health and Care Excellence (NICE) recommends between 7 and 10 visits for uncomplicated pregnancies for parous and nulliparous women, respectively [94]. In addition, NICE provides clear recommendations for the assessment and monitoring of possible depression during pregnancy [95]. In the first instance, clinicians should consider using the Whooley questions to initiate a general discussion about a woman's mental health and wellbeing. If a patient responds affirmatively to any of these questions, the Edinburgh Postnatal Depression Scale (EPDS) or the Patient Health Questionnaire (PHQ-9) should be used as part of a complete clinical assessment. Alternatively, the patient should be referred to her general practitioner (GP) or to a mental health professional, depending on the suspected severity. In addition, this guideline recommends using the EPDS or the PHQ-9 for the ongoing monitoring of depression [95]. Formal assessment of depression is recommended as part of the booking appointment [94, 95]. This is the first official antenatal appointment and it usually takes place at around week

ten. In addition, women should be asked about their emotional wellbeing at each subsequent antenatal appointment, although no formal assessment is currently recommended.

The highly-structured nature of antenatal care offers a unique opportunity to evaluate the potential role of smartphones and tablet computer apps for implementing the recommendations of these guidelines.

1.5. Thesis overview

1.5.1. Aim

The aim of this work is to evaluate the role of apps running on smartphones and tablet computers for the assessment and monitoring of psychological wellbeing and depressive symptoms during pregnancy through self-administered surveys, in the context of antenatal care provided through the NHS in England.

1.5.2. Chapter objectives

Through a systematic review, Chapter 2 assesses the impact that smartphone and tablet computer apps as a delivery mode have on the quality of self-administered survey responses, compared to traditional delivery modes. This chapter also discusses the role that the setting in which surveys are completed has in informing decisions about data collection strategies and quality assurance measures, as well as the factors that could lead to delivery mode effects.

Chapter 3 explores the feasibility of using a commercial, general-purpose app running on tablet computers to administer the Whooley questions and the EPDS to pregnant women in the waiting area of NHS antenatal clinics in England. In doing so, this chapter discusses the effect that survey layout (i.e., vertical scrolling versus pagination) has on EPDS responses, as well as the implementation challenges associated with this approach to mental health assessment in antenatal care.

Chapter 4 explores how a tablet computer app was used to evaluate the diagnostic properties of the Whooley questions in relation to EPDS scores, and outlines the role that this technology has in maximising the scalability of diagnostic accuracy studies.

Chapter 5 explores the feasibility of using a bespoke app, BrightSelf, running on pregnant women's own devices to monitor psychological wellbeing and symptoms of depression during the last 2 trimesters of pregnancy. In doing so, this chapter explores notions of adherence and the suitability of momentary questions as a strategy for promoting effective patient engagement.

Chapter 6 provides an overall discussion of the findings of this project, particularly in relation to the methodological, attitudinal and practical concerns outlined in this introductory chapter.

References (Chapter 1)

1. Health Sciences Library, University of Washington – University Libraries. *Data Resources in the Health Sciences*. Available from: http://guides.lib.uw.edu/hsl/data/findclin [Accessed 19 January 2017].
2. Donaldson LJ, Scally G (eds.). *Donaldsons' Essential Public Health*. 3rd Edition. Oxford, UK: Radcliffe Publishing; 2009.
3. British Medical Association. *Secondary uses of data*. Available from: https://www.bma.org.uk/collective-voice/policy-and-research/ethics/secondary-uses-of-data [Accessed 19 January 2017].
4. NHS Digital. *Data sets*. Available from: http://content.digital.nhs.uk/datasets [Accessed 19 January 2017].
5. Yong PL, Olsen LA, McGinnis JM (eds.). *Value in Health Care: Accounting for Cost, Quality, Safety, Outcomes and Innovation*. Washington (DC): National Academies Press (US); 2010. Available from: https://www.ncbi.nlm.nih.gov/books/NBK50934/ [Accessed 19 January 2017].
6. Hosking JD, Newhouse MM, Bagniewska A, Hawkins BS. Data collection and transcription. *Controlled Clinical Trials*. 1995;16(2 Suppl): 66S-103S. Available from: doi: http://dx.doi.org/10.1016/0197-2456(94)00094-J [Accessed 19 January 2017].
7. Bowling, A. Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health (Oxford, England)*. 2005;27(3): 281-91. Available from: doi: 10.1093/pubmed/fdi031 [Accessed 19 of January 2017].
8. Wilcox AB, Gallagher KD, Boden-Albala B, Bakken SR. Research data collection methods: from paper to tablet computers. *Medical Care*. 2012;50(Suppl): S68-73. Available from: doi: 10.1097/MLR.0b013e318259c1e7 [Accessed 19 January 2017].
9. Boynton P, Greenhalgh T. Hands-on guide to questionnaire research: Selecting, designing, and developing your questionnaire. *BMJ*. 2004;328: 1312-1315. Available from: doi: https://doi.org/10.1136/bmj.328.7451.1312 [Accessed 20 January 2017].
10. Haller G, Haller DM, Courvoisier DS, Lovis C. Handheld vs. Laptop Computers for Electronic Data Collection in Clinical Research: A Crossover Randomized Trial. *Journal of the American Medical Informatics Association (JAMIA)*. 2009;16(5): 651-9. Available from: doi: 10.1197/jamia.M3041 [Accessed 20 January 2017].

11. Lee YW, Strong DM, Kahn BK, Wang RY. AIMQ: a methodology for information quality assessment. *Information & Management*. 2002;40(2): 133-146. Available from: doi: 10.1016/S0378-7206(02)00043-5 [Accessed 20 January 2017].
12. Pipino LL, Lee YW, Yang RY. Data Quality Assessment. *Communications of the ACM*. 2002;45(4): 211-18. Available from: http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf [Accessed 20 January 2017].
13. Van den Eynden V, Corti L, Woollard M, Bishop L, Horton L. *Managing and Sharing Data*. 3rd Edition. Essex, UK: UK Data Archive, University of Essex; 2011.
14. Tai B, Boyle M, Ghitza U, Kaplan RM, Clark HW, Gersing K. Meaningful use of electronic behavioral health data in primary health care. *Science Translational Medicine*. 2012;4(119): 119mr3. Available from: doi: 10.1126/scitranslmed.3003324 [Accessed 20 January 2017].
15. Goben A, Raszewski R. The data life cycle applied to our own data. *Journal of the Medical Library Association*. 2015;103(1): 40-4. Available from: doi: 10.3163/1536-5050.103.1.008 [Accessed 20 January 2017].
16. UK Data Service. *Research Data Lifecycle*. Available from: https://www.ukdataservice.ac.uk/manage-data/lifecycle [Accessed 25 July 2018].
17. Lampe AJ, Weiler JM. Data Capture From the Sponsors' and Investigators' Perspectives: Balancing Quality, Speed, and Cost. *Therapeutic Innovation & Regulatory Science*. 1998;32(4): 871-86. Available from: doi: https://doi.org/10.1177/009286159803200403 [Accessed 21 January 2017].
18. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, Lenderking WR, Cella D, Basch E, ISPOR ePRO Task Force. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. *Value in Health*. 2009;12(4): 419-29. Available from: doi: 10.1111/j.1524-4733.2008.00470.x [Accessed 20 January 2017].
19. University of Minnesota. *Data Collection Techniques.* Available from: https://cyfar.org/data-collection-techniques [Accessed 28 July 2018].
20. Bowling A. *Research Methods in Health: Investigating Health and Health Services*. 3rd Edition. Maidenhead, Berkshire: McGraw Hill; 2009.

21. Carter Y, Shaw S, Thomas C (eds.) *The Use and Design of Questionnaires*. Master Classes in Primary Care Research No 2. London, UK: The Royal College of General Practitioners; 2000.
22. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey Methodology*. 2nd Edition. New Jersey, US: Wiley; 2009.
23. Lavrakas, PJ (ed.). *Encyclopedia of Survey Research Methods*. London, UK: SAGE Publications; 2008.
24. Groves RM, Lyberg L. Total Survey Error: Past, Present, Future. *Public Opinion Quarterly*. 2010;74(5):849-879. Available from: doi: 10.1093/poq/nfq065 [Accessed 28 July 2018].
25. Dowrick AS, Wootten AC, Murphy DG, Costello AJ. "We Used a Validated Questionnaire": What Does This Mean and Is It an Accurate Statement in Urologic Research? *Health Services Research*. 2015;85(6):1304-1311. Available from: doi: 10.1016/j.urology.2015.01.046 [Accessed 28 July 2018].
26. Tourangeau R, Rips LJ, Rasinski K. *The Psychology of Survey Response*. New York, US: Cambridge University Press; 2000.
27. Fan W, Yan Z. Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*. 2010;26(2): 132-9. Available from: doi: 10.1016/j.chb.2009.10.015 [Accessed 22 January 2017].
28. Wells T, Bailey JT, Link MW. Comparison of Smartphone and Online Computer Survey Administration. *Social Science Computer Review*. 2014;32(2): 238-255. Available from: doi: 10.1177/0894439313505829 [Accessed 22 January 2017].
29. World Health Organization. *mHealth: New horizons for health through mobile technologies*. Global Observatory for eHealth series – Volume 3. Geneva, Switzerland: World Health Organization; 2011.
30. Tomlinson M, Rotheram-Borus MJ, Swartz L, Tsai AC. Scaling Up mHealth: where is the evidence? *PLoS Med*. 2013;10(2): e1001382. Available from: doi: https://doi.org/10.1371/journal.pmed.1001382 [Accessed 21 January 2017].
31. Dobkin BH, Dorsch A. The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. *Neurorehabilitation and Neural Repair*. 2011;25(9): 788-98. Available from: doi: 10.1177/1545968311425908 [Accessed 20 January 2017].

32. Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*. 2012;45(1): 184-98. Available from: doi: 10.1016/j.jbi.2011.08.017 [Accessed 20 January 2017].
33. Trucano M. *Using mobile phones in data collection: Opportunities, issues and challenges*. Available from: http://blogs.worldbank.org/edutech/using-mobile-phones-data-collection-opportunities-issues-and-challenges [Accessed 20 January 2017].
34. Gaggioli A, Pioggia G, Tartarisco G, Baldus G, Corda D, Cipresso P, Riva G. A mobile data collection platform for mental health research. *Personal and Ubiquitous Computing*. 2013;17(2): 241-251. Available from: doi: 10.1007/s00779-011-0465-2 [Accessed 22 January 2017].
35. Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG. EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. *PLoS One*. 2009;4(9): e6968. Available from: doi: 10.1371/journal.pone.0006968 [Accessed 20 January 2017].
36. Aanensen DM, Huntley DM, Menegazzo M, Powell CI Spratt BG. EpiCollect+: linking smartphones to web applications for complex data collection projects. *F1000Research*. 2014;3: 199. Available from: doi: 10.12688/f1000research.4702.1 [Accessed 20 January 2017].
37. Apple®. *iTunes Preview*. Available from: https://itunes.apple.com/gb/genre/ios/id36?mt=8 [Accessed 23 January 2017].
38. Google ©. *Google Play*. Available from: https://play.google.com/store?hl=en_GB [Accessed 23 January 2017].
39. Ofcom. *Communications Market Report 2017*. Ofcom, 2017.
40. Deloitte. *Global Mobile Consumer Survey 2016: UK Cut*. Deloitte, 2016.
41. Oulasvirta A, Rattenbury T, Ma L, Raita E. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing* 2012;16(1): 105-114. Available from: doi: 10.1007/s00779-011-0412-2 [Accessed 23 January 2017].
42. Andrews S, Ellis DA, Shaw H, Piwek L. Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use. *PLOS One*. 2015;10(10): e0139004. Available from: doi: 10.1371/journal.pone.0139004 [Accessed 23 January 2017].

43. Yildirim C, Correia AP. Exploring the dimensions of nomophobia: Development and validation of a self-reported questionnaire. *Computers in Human Behavior*. 2015;49: 130-7. Available from: doi: 10.1016/j.chb.2015.02.059 [Accessed 23 January 2017].
44. Chittaranjan G, Blom J, Gatica-Perez D. Mining large--scale smartphone data for personality studies. *Personal and Ubiquitous Computing*. 2013;17(3): 433-450. Available from: doi: 10.1007/s00779-011-0490-1 [Accessed 23 January 2017].
45. Krishna S, Boren SA, Balas EA. Healthcare via cell phones: a systematic review. *Telemedicine Journal and eHealth*. 2009;15(3): 231-240. Available from: doi: 10.1089/tmj.2008.0099 [Accessed 23 January 2017].
46. Ofcom. *Access and Inclusion in 2016*. Ofcom, 2017.
47. Sugie NF. Utilizing Smartphones to Study Disadvantaged and Hard-to-Reach Groups. *Sociological Methods and Research*. 2018;47(3):458-491. Available from: doi: 10.1177/0049124115626176 [Accessed 25 July 2018].
48. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient compliance with paper and electronic diaries. *Controlled Clinical Trials*. 2003;24(2): 182-199.
49. Marcano Belisario JS, Huckvale K, Greenfield G, Car J, Gunn L. Smartphone and tablet self management apps for asthma. *Cochrane Database of Systematic Reviews.* 2013;27(11): CD010013. Available from: doi: 10.1002/14651858.CD010013.pub2 [Accessed 24 January 2017].
50. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. *The International Journal of Behavioral Nutrition and Physical Activity*. 2014;11: 97. Available from: doi: 10.1186/s12966-014-0097-9 [Accessed 24 January 2017].
51. BinDhim NF, Alanazi EM, Aljadhey H, Basyouni MH, Kowalski SR, Pont LG, Shaman AM, Trevena L, Alhawassi TM. Does a Mobile Phone Depression-Screening App Motivate Mobile Phone Users With High Depressive Symptoms to Seek a Health Care Professional's Help? *Journal of Medical Internet Research*. 2016;18(6): e156. Available from: doi: 10.2196/jmir.5726 [Accessed 24 January 2017].
52. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence Based Mental Health*. 2018. Available from: doi: 10.1136/eb-2018-102891 [Accessed 25 July 2018].

53. Runyan JD, Steenbergh TA, Brainbridge C, Daugherty DA, Oke L, Fry BN. A smartphone ecological momentary assessment/intervention "app" for collecting real-time data and promoting self-awareness. *PLoS One*. 2013;8(8): e71325. Available from: doi: 10.1371/journal.pone.0071325 [Accessed 24 January 2017].
54. Runyan JD, Steinke EG. Virtues, ecological momentary assessment/intervention and smartphone technology. *Frontiers in Psychology*. 2015;6: 481. Available from: doi: 10.3389/fpsyg.2015.00481 [Accessed 24 January 2017].
55. Kahneman D. *Thinking, Fast and Slow*. London, UK: Penguin; 2012.
56. Ariely D. *Predictably Irrational*. London, UK: HaperCollins; 2009.
57. Arsoniadis EG, Tambyraja R, Khairat S, Jahansouz C, Scheppmann D, Kwaan MR, Hultman G, Melton GB. Characterizing Patient-Generated Clinical Data and Associated Implications for Electronic Health Records. *Studies in Health Technology and Informatics*. 2015;216: 158-162.
58. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, Merchant GC, Naughton F, Blandford A. Understanding and Promoting Effective Engagement With Digital Behaviour Change Interventions. *American Journal of Preventive Medicine*. 2016;51(5):833-842. Available from: doi: 10.1016/j.amepre.2016.06.015 [Accessed 26 July 2018].
59. Torous J, Chan SR, Tan SYM, Behrens J, Mathew I, Conrad EJ, Hinton L, Yellowlees P, Keshavan M. Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics. *JMIR Mental Health*. 2014;1(1): e5. Available from: doi: 10.2196/mental.4004 [Accessed 25 January 2017].
60. Patrick K, Hekler EB, Estrin D, Mohr DC, Riper H, Crane D, Godino J, Riley WT. The Pace of Technologic Change: Implications for Digital Health Behavior Intervention Research. *American Journal of Preventive Medicine*. 2016;51(5):816-824. Available from: doi: 10.1016/j.amepre.2016.05.001 [Accessed 26 July 2018].
61. Abroms LC, Lee WJ, Bontemps-Jones J, Ramani R, Mellerson J. A content analysis of popular smartphone apps for smoking cessation. *American Journal of Preventive Medicine.* 2013;45(6): 732-6. Available from: doi: 10.1016/j.amepre.2013.07.008 [Accessed 25 January 2017].
62. Ubhi HK, Kotz D, Michie S, van Schayck OC, Sheard D, Selladurai A, West R. Comparative analysis of smoking cessation smartphone applications available in 2012 versus 2014.

Addict Behaviors. 2016;58: 175-181. Available from: doi: 10.1016/j.addbeh.2016.02.026 [Accessed 25 January 2017].

63. Morrissey EC, Corbett TK, Walsh JC, Molloy GJ. Behavior Change Techniques in Apps for Medication Adherence: A Content Analysis. *American Journal of Preventive Medicine*. 2016;50(5): e143-6. Available from: doi: 10.1016/j.amepre.2015.09.034 [Accessed 25 January 2017].
64. Crane D, Garnett C, Brown J, West R, Michie S. Behavior change techniques in popular alcohol reduction apps: content analysis. *Journal of Medical Internet Research*. 2015; 17(5): e118. Available from: doi: 10.2196/jmir.4060 [Accessed 25 January 2017].
65. Conroy DE, Yang CH, Maher JP. Behavior change techniques in top-ranked mobile apps for physical activity. *Americal Journal of Preventive Medicine*. 2014;46(6): 649-652. Available from: doi: 10.1016/j.amepre.2014.01.010 [Accessed 25 January 2017].
66. Muessig KE, Pike EC, Legrand S, Hightow-Weidman LB. Mobile phone applications for the care and prevention of HIV and other sexually transmitted diseases: a review. *Journal of Medical Internet Research*. 2013;15(1): e1. Available from: doi: 10.2196/jmir.2301 [Accessed 25 January 2017].
67. BinDhim NF, Shaman AM, Trevena L, Basyouni MH, Pont LG, Alhawassi TM. Depression screening via a smartphone app: cross-country user characteristics and feasibility. *Journal of the American Medical Informatics Association (JAMIA).* 2015;22(1): 29-34. Available from: doi: 10.1136/amiajnl-2014-002840 [Accessed 25 January 2017].
68. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker Mobile Web App With Email for Adults to Self-Manage Mild-to-Moderate Depression: Randomized Controlled Trial. *JMIR Mhealth and Uhealth*. 2016;4(1): e8. Available from: doi: 10.2196/mhealth.4231 [Acessed 25 January 2017].
69. Faurholt-Jepsen M, Frost M, Vinberg M, Christensen EM, Bardram JE, Kessing LV. Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry Research*. 2014;217(1-2): 124-7. Available from: doi: 10.1016/j.psychres.2014.03.009 [Accessed 25 January 2017].
70. Tabb KM, Choi S, Pineros-Leano M, Meline B, McDonald HG, Kester R, Huang H. Perinatal depression screening in a Women, Infants, and Children (WIC) program: perception of feasibility and acceptability among a multidisciplinary staff. *Gen Hospital Psychiatry*.

2015;37(4): 305-9. Available from: doi: 10.1016/j.genhosppsych.2015.03.008 [Accessed 25 January 2017].

71. Mental Health Foundation. *Fundamental Facts About Mental Health 2015*. Available from: https://www.mentalhealth.org.uk/publications/fundamental-facts-about-mental-health-2015 [Accessed 25 January 2017].
72. Royal College of Psychiatrists. Physical Health in Mental Health. Available from: http://www.rcpsych.ac.uk/usefulresources/publications/collegereports/op/op67.aspx [Accessed 25 January 2017].
73. Levis B, Benedetti A, Riehm KE, Saadat N, Levis AW, Azar M, et al. Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. *The British Journal of Psychiatry*. 2018;212: 377-385. Available from: doi: 10.1192/bjp.2018.54 [Accessed 26 July 2018].
74. Moskowitz DS, Young SN. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience*. 2006;31(1): 13-20.
75. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges. *Journal of Medical Internet Research*. 2017;19(3):e77. Available from: doi: 10.2196/jmir.7138 [Accessed 26 July 2018].
76. Stewart DE. Clinical practice. Depression during pregnancy. *New England Journal of Medicine*. 2011;365(17): 1605-11. Available from: doi: 10.1056/NEJMcp1102730 [Accessed 25 January 2017].
77. Thombs BD, Stewart DE. Depression screening in pregnancy and postpartum: how close are we? *Journal of Psychosomatic Research*. 2014;77(3): 244-5. Available from: doi: 10.1016/j.jpsychores.2014.07.017 [Accessed 25 January 2017].
78. Thombs BD, Stewart DE. Depression screening in pregnancy and postpartum: who needs evidence? *Journal of Psychosomatic Research*. 2014;76(6): 492-3. Available from: doi: 10.1016/j.jpsychores.2014.03.102 [Accessed 25 January 2017].
79. Gavin NI, Gaynes BN, Lohr KN, Meltzer-Brody S, Gartlehner G, Swinson T. Perinatal depression: a systematic review of prevalence and incidence. *Obstetrics and Gynecology*. 2005; 106(5 Pt 1): 1071-83.

80. Lee AM, Lam SK, Sze Mun Lau SM, Chong CS, Chui HW, Fong DY. Prevalence, course, and risk factors for antenatal anxiety and depression. *Obstetrics and Gynecology.* 2007; 110(5): 1102-12.
81. Dagklis T, Papazisis G, Tsakiridis I, Chouliara F, Mamopoulos A, Rousso D. Prevalence of antenatal depression and associated factors among pregnant women hospitalized in a high-risk pregnancy unit in Greece. *Social Psychiatry Psychiatric Epidemiology*. 2016;51(7): 1025-31. Available from: doi: 10.1007/s00127-016-1230-7 [Accessed 25 January 2017].
82. Bennett HA, Einarson A, Taddio A, Koren G, Einarson TG. Prevalence of depression during pregnancy: systematic review. *Obstetrics and Gynecology*. 2004;103(4): 698-709.
83. Parker GB, Hegarty B, Paterson A, Hadzi-Pavlovic D, Granville-Smith I, Gokiert A. Predictors of post-natal depression are shaped distinctly by the measure of 'depression'. *Journal of Affective Disorders*. 2015;173: 239-44. Available from: doi: 10.1016/j.jad.2014.10.066 [Accessed 25 January 2017].
84. Heron J, O'Connor TG, Evans J, Golding J, Glover V, ALSPAC Study Team. The course of anxiety and depression through pregnancy and the postpartum in a community sample. *Journal of Affective Disorders*. 2004;80(1): 65-73.
85. Accortt EE, Cheadle AC, Dunkel Schetter C. Prenatal depression and adverse birth outcomes: an updated systematic review. *Maternal and Child Health Journal.* 2015;19(6): 1306-37. Available from: doi: 10.1007/s10995-014-1637-2 [Accessed 25 January 2017].
86. Chen CH, Lin HC. Prenatal care and adverse pregnancy outcomes among women with depression: a nationwide population-based study. *Canadian Journal of Psychiatry*. 2011; 56(5): 273-80.
87. Kinsella MT, Monk C. Impact of maternal stress, depression and anxiety on fetal neurobehavioral development. *Clinical Obstetrics and Gynecology*. 2009;52(3): 425-40. Available from: doi: 10.1097/GRF.0b013e3181b52df1 [Accessed 25 January 2017].
88. Pearson RM, Evans J, Kounali D, Lewis G, Heron J, Ramchandani PG, O'Connor TG, Stein A. Maternal depression during pregnancy and the postnatal period: risks and possible mechanisms for offspring depression at age 18 years. *JAMA Psychiatry*. 2013;70(12): 1312-9. Available from: doi: 10.1001/jamapsychiatry.2013.2163 [Accessed 25 January 2017].

89. Braithwaite EC, Murphy SE, Ramchandani PG. Effects of prenatal depressive symptoms on maternal and infant cortisol reactivity. *Archives of Women's Mental Health*. 2016; 19(4): 581-90. Available from: doi: 10.1007/s00737-016-0611-y [Accessed 25 January 2017].
90. Bauer A, Parsonage M, Knapp M, Iemmi V, Adelaja B. The cost of perinatal mental health problems. London, UK: Centre for Mental Health and London School of Economics; 2014.
91. Howard LM, Molyneaux E, Dennis CL, Rochat T, Stein A, Milgrom J. Non-psychotic mental disorders in the perinatal period. *Lancet*. 2014;384(9956): 1775-88. Available from: doi: 10.1016/S0140-6736(14)61276-9 [Accessed 26 January 2017].
92. Kingston D, Austin MP, Heaman M, McDonald S, Lasiuk G, Sword W, Giallo R, Hegadoren K, Vermeyden L, van Zanten SV, Kingston J, Jarema K, Biringer A. Barriers and facilitators of mental health screening in pregnancy. *Journal of Affective Disorders*. 2015;186: 350-7. Available from: doi: 10.1016/j.jad.2015.06.029 [Accessed 26 January 2017].
93. Thomas N, Komiti A, Judd F. Pilot early intervention antenatal group program for pregnant women with anxiety and depression. *Archives of Women's Mental Health*. 2014; 17(6): 503-9. Available from: doi: 10.1007/s00737-014-0447-2 [Accessed 26 January 2017].
94. National Institute for Health and Clinical Excellence (2008 – 2017 Update). *Antenatal care for uncomplicated pregnancies*. Clinical Guideline CG62.
95. National Collaborating Centre for Mental Health (2014 – 2015 Update), *Antenatal and postnatal mental health: clinical management and service guidance*. Clinical Guideline CG192.

2. Effect of apps as a delivery mode on the quality of self-administered survey responses: a systematic review

2.1. Overview

The quantitative survey method is one of the most widely used data collection methods in healthcare. Surveys represent the cornerstone of this method, as they enable the collection of vast amounts of highly structured data that are standardised across patients and sites. Mobile apps have the potential to further leverage the advantages of the survey quantitative method by minimising the data management requirements of surveys. In addition, apps could enable the distribution of surveys to an increased number of geographically dispersed individuals. However, apps also have the potential to influence the properties of survey responses, independently from the clinical constructs that are being assessed. This phenomenon is known as delivery mode effect, and it could have important implications for diagnostic and therapeutic decisions.

This chapter presents a systematic review that assessed the impact that apps can have on the quality of responses given to self-administered health surveys, when compared to other delivery modes. Data quality was interpreted in the context of the measurement error framework, and the following indicators were chosen: data equivalence, data accuracy, time needed to complete a survey, response rates, data completeness, adherence to survey administration protocols, and acceptability to patients.

Fourteen studies were included in this systematic review. The findings showed that responses collected through apps are equivalent to those collected through other delivery modes. This appears to be the case when the intended clinical use of a survey, its intended frequency of administration, and the conditions in which it was validated resemble those of the project in question. There was no data available on data accuracy or response rates. The findings concerning the time needed to complete a survey were inconclusive. Moreover, although apps might improve data completeness, there was insufficient evidence to assess their impact on adherence to survey completion schedules. Acceptability to respondents was shown to be a useful construct to identify usability issues that might influence the successful adoption of apps in clinical contexts. This systematic review also identified key factors that could result in

delivery mode effects when using apps: survey design, app functionality, human-computer interaction factors, characteristics of the patients or of their clinical condition, and survey completion setting.

These findings informed subsequent work that evaluated the use of mobile apps for the identification and monitoring of depression during pregnancy in the context of antenatal care provided by the NHS in England.

2.2. Rationale

The quantitative survey method is one of the most commonly used data collection methods in healthcare, as it enables the collection of highly structured data that are standardised across patients (or respondents) and across sites [1-5]. This method has become the basis of evidence for many clinical decisions, public health policy, and the development and evaluation of interventions. However, the mode through which surveys are delivered could affect the quality of the responses collected [1, 6]. Apps running on consumer mobile devices can operate as a delivery mode for health-related surveys. Therefore, the impact that this new delivery mode could have on the quality of survey responses needs careful consideration.

2.3. Background

Surveys are the key tool of the quantitative survey method. They typically consist of a series of items (i.e., questions or statements) designed to gather information from a sample of individuals [6]. As such, surveys can enable the collection of large amounts of highly structured data that are standardised across patients and sites [1-5]. Moreover, the popularity of these instruments can be partly attributed to the increased awareness amongst clinicians of the impact that healthcare interventions can have on the quality of a patient's life, and how this can influence a patient's ability to cope with illness or to recover from it [7]. In addition, the effectiveness of therapeutic efforts in many medical disciplines and allied healthcare professions is driven primarily by improvements in quality of life [7]. Furthermore, many of the phenomena that these disciplines are interested in (e.g., mood, pain, functionality) tend to be introspective and experiential in nature. Therefore, the assessment of this wide range of phenomena typically lies in the realm of surveys (as opposed to traditional lab-based measurements).

According to Bowling [6], surveys can be of two types: descriptive or analytic. The aim of descriptive surveys is to measure certain attributes in a target population. A sample of respondents are asked to report on the attribute(s) of interest (i.e., events, behaviours, attitudes, opinions, feelings, and knowledge), usually at one point in time and retrospectively. The responses collected are used to estimate relevant statistics on individuals and on the sample of respondents, which are then used to estimate relevant parameters on the target

population. The aim of analytic surveys on the other hand, is to investigate associations between variables [6]. With these surveys, respondents are asked to report on the attribute(s) of interest at more than one point in time (i.e., longitudinally). Depending on the frequency and duration of the sampling protocol, these surveys can be used to identify the direction of the association, and the presence of seasonal effects [6].

Surveys can also differ in terms of data collection mode. This refers to variation in any stage of the survey process: (i) the method used to initiate contact with potential respondents; (ii) the medium used to deliver the survey to respondents; and (iii) the medium used to administer the actual questions [1, 8]. Of these, the latter is thought to exert great influence on survey responses [1].

In relation to the administration of questions, surveys can be interview-administered or self-administered. While both approaches have their advantages and disadvantages (Table 1), self-administration is usually preferred. Self-administered surveys are ideal for reaching relatively large and geographically-dispersed populations [1, 3, 6, 9]. In addition, they are better suited to deal with sensitive topics, and are less resource-intensive than interview-administered surveys [3, 6]. To further leverage these benefits whilst reducing costs, the electronic delivery of self-administered surveys has received special attention [4, 10-13]. For example, electronic surveys have been used in areas such as pain, asthma, tobacco use, and smoking cessation [12].

However, survey responses are the result of a complex interaction between surveys, respondents and the delivery mode [1, 3, 10, 14-16]. Electronic delivery modes vary in the type and degree of technology used to deliver surveys, the channels through which questions and response options are presented to respondents, and the data entry formats that are supported [1, 4, 9, 15]. Variation in any of these domains may affect the properties of responses, a phenomenon known as delivery mode effect [1, 10, 16, 17].

Table 1. Advantages and disadvantages of interview-administered and self-administered surveys (adapted from Carter Y, Shaw S, Thomas C, p. 13)

			Interview-administered	Self-administered
Response rates	General population samples		Best	Poor to good
	Special population samples		Good to best	Satisfactory to good
Representative samples	Avoidance of refusal bias		Dependent upon interviewer's technique	Poor
	Control over who completes the survey		Moderate to good	Poor to good
Ability to handle:	Long surveys		Moderate to good	Poor to satisfactory
	Complex and/or boring questions		Moderate to good	Poor to moderate
	Item non-response		Good	Moderate
	Question sequence control		Good	Poor
	Open-ended questions		Good	Poor
Quality of responses	Minimise social desirability bias		Poor to moderate	Satisfactory
	Ability to avoid distortion due to:	Interviewer's characteristics	Poor to moderate	Good
		Interviewer's opinion	Moderate	Good
		Other people	Moderate to good	Poor
	Allow opportunities to consult		Poor to moderate	Good
Survey implementation	Ease of identifying suitable staff		Poor to moderate	Good
	Speed		Poor to good	Poor
	Cost		Poor to moderate	Good

Delivery mode effects can be conceptualised as a data quality dimension. In this context, quality is understood in terms of survey error: errors of measurement and errors of representation [4, 8]. The former refers to any discrepancies between the responses given to a survey and the true value of the attribute of interest [4, 7, 8]. This type of error could manifest itself as differences in the statistics estimates, or as several types of response biases (i.e., social desirability bias, acquiescence or extremeness bias, recall effects or response order effects) [1, 4]. For example, survey responses collected electronically have been found to be more accurate and timely than, or (in the case of validated instruments) equivalent to, those obtained using paper surveys [9, 12]. Moreover, there is no clear evidence of potential recall or social desirability biases when using electronic delivery modes [1]. However, primacy effects have been found to be more prevalent when response options are presented visually, while recency effects are more common when response options are presented aurally [4, 8].

Errors of representation refer to any discrepancies between the estimates made on a sample of respondents, and those on the target population [8]. This type of error is best understood in terms of the effect that a delivery mode has on sampling error, coverage error, and non-response error [8]. The evidence suggests that, compared to other delivery modes, web surveys can decrease response rates by 10% to 20% [13, 18]. However, this effect is mediated by the content and presentation of the survey, the sampling methods used, the type and number of invitations sent to potential respondents, access to technology, and the stage of the survey questionnaire process [1, 13, 14, 16]. Moreover, adherence to sampling protocols tends to increase with electronic delivery modes [9, 12]. In addition, electronic surveys result in higher item response rates than paper surveys [1].

Delivery mode effects tend to be mode-specific; therefore, it cannot be assumed that lessons from one mode will apply to all others [16]. Recent development in mobile technology have produced new types of consumer devices (i.e., smartphones and tablet computers) that can operate as a delivery mode for surveys. Delivery mode effects have been documented for a wide range of electronic devices: laptop and desktop computers, mobile phones (feature phones), PDAs, and web surveys [1, 9, 10, 12, 14]. However, they have not been systematically assessed for consumer mobile devices.

2.3.1. Apps and consumer mobile devices

Smartphones and tablet computers are mobile devices with advanced computing, sensing and networking capabilities, and with an OS framework that focuses on apps. These consumer devices typically operate as general purpose devices, enabling users to perform a wide range of functions: browsing the internet, managing calendars or taking photos. However, through the OS, apps can also enable smartphones or tablet computers to perform specialist functions. In this case, a tool for the administration of health surveys.

Compared to other electronic delivery modes, consumer mobile devices offer the perceived advantage of being able to conveniently complete surveys at any time and anywhere, as they are almost always on and on a person. Survey completion could be further enhanced by app interfaces aimed at maximising the quality of survey responses. For example, alerts and reminders could increase data completeness, and the intermittent presentation of questions

could make survey completion more compatible with the typical usage pattern of these devices. Moreover, the connectivity, portability and ubiquity of consumer mobile devices could allow the completion of surveys in a multitude of settings and situational contexts. These characteristics could influence the information available to respondents, the salience of certain cues, the speed required to produce responses, the chosen accuracy for responses, and the influence or social norms operating at that particular time [15]. These influences will determine the specific cognitive processes required to generate a response, and will thus affect the final properties of the survey responses [15, 19, 20].

Therefore, assessing potential delivery mode effects on health survey responses collected through apps is warranted for the following reasons:

- Lack of a comprehensive assessment of this new delivery mode
- Importance of self-administered surveys in gathering information, and generating the evidence, for a wide range of healthcare activities
- Number of practitioners and researchers already using apps for delivering self-administered surveys

The purpose of this chapter is to provide such assessment.

2.4. Methods

The methods implemented in this chapter followed the recommendations of the Cochrane Collaboration, and were published as a Cochrane systematic review [17].

2.4.1. Aim

The aim of this systematic review was to assess the impact that smartphone and tablet computer apps have on the quality of self-administered survey responses, compared to alternative delivery modes.

2.4.2. Inclusion criteria

2.4.2.1. Types of studies

Following the recommendations of The International Society for Pharmacoeconomics and Outcomes Research electronic Patient-Reported Outcome (ISPOR ePRO) Good Research Practices Task Force Report [10], the following types of studies were included:

- Parallel, randomised controlled trials (RCT)
- Crossover trials
- Paired repeated measures study designs

No exclusions were made on the basis of type of publication. However, risks associated with the inclusion of non-peer reviewed reports (i.e., conference abstracts and posters) were considered during the risk of bias assessments.

2.4.2.2. Types of data

To be included, data had to be obtained from individuals completing health surveys, either validated or non-validated. Participants could be healthy volunteers or patients with any clinical diagnosis. In addition, the original intention was to include data from patients who were completing self-administered surveys as part of complex self-management interventions. However, no such studies were identified.

There was no exclusion based on the age, gender or socio-demographic background of the respondents.

Data from the following sources were excluded: interviewers, clinicians, caregivers or parents completing surveys on behalf of someone else; surveys measuring consumer behaviour; and surveys used as part of routine paperwork.

2.4.2.3. Types of methods

To be included, studies had to use a smartphone or tablet computer app for delivering self-administered surveys. Only native apps (i.e., those that had been developed for a specific mobile platform, such as Android or iPhone Operating System [iOS]), or web apps wrapped

within a native app framework (such as Adobe® PhoneGap™) were considered. Web apps rendered on a web browser were excluded, as they differ from native apps in terms of usability (e.g., responsiveness) and performance.

Only smartphones and tablet computers that became available in 2007 or later were included, as these devices are compatible with the current mobile OS that focuses on apps.

Included studies had to compare at least two delivery modes, one of which had to be a smartphone or tablet computer app.

Studies assessing pictures as a data entry format were excluded, as well as studies in which students, researchers or employees used smartphones and tablet computer apps to collect data as part of their regular duties.

2.4.2.4. Types of outcomes

The choice of outcomes was informed by a review of the medical literature, which identified the data quality domains relevant to healthcare.

2.4.2.4.1. Primary outcomes

- Equivalence between survey responses administered via two different delivery modes: in this context, data equivalence refers to changes in the psychometric properties of a survey when it is adapted for use with a new delivery mode. This outcome was measured using the following indicators:
    - Correlation coefficients or measures of agreement (intra-class correlation [ICC] coefficients, Pearson product-moment correlations, Spearman rho and weighted Kappa coefficients)
    - Comparison of mean scores between two delivery modes
    - Both types of measures

  The focus was on a survey's overall equivalence, as opposed to equivalence between constructs or individual items [9]. For ICC coefficients, a cut-off of 0.70 was used for group comparisons; for other coefficients, a cut-off of 0.60 was used [9]. In studies

comparing mean scores, equivalence was assessed in relation to the survey's minimally important difference (MID). Whenever possible, between-mode mean differences and ICC coefficients were interpreted in relation to the within-mode test-retest ICC of the original delivery mode [9, 10]

- Data accuracy: defined as the comparison of the proportion of errors or problematic items between two different delivery modes
- Data completeness: defined as the comparison of the proportion of missing items between two different delivery modes
- Response rates: defined as the number of completed questionnaires as a proportion of the total number of eligible sample units

2.4.2.4.2. Secondary outcomes

- Time needed to complete a survey
- Respondents' adherence to a pre-specified sampling protocol (both in terms of frequency and duration of the assessments)
- Acceptability of the delivery mode to respondents

2.4.3. Search methods

The following electronic bibliographic databases were searched: MEDLINE; EMBASE; PscyINFO; IEEEXplore; Web of Science; CABI: CAB Abstracts; Current Contents Connect; ACM Digital; ERIC; Sociological Abstracts; Health Management Information Consortium; the Campbell Library; and CENTRAL (see Appendices 1 to 12). In addition, the following registers of current and ongoing trials were searched: ClinicalTrials.gov (Appendix 13) and the World Health Organization (WHO) International Clinical Trials Registry Platform. Repositories of grey literature were also searched: OpenGrey; Mobile Active; and ProQuest Dissertation & Theses (Appendix 14). In addition, the reference lists of potentially relevant studies were also examined, as well as the public profiles of authors with relevant publications.

All databases were searched for studies published between January 2007 and April 2015, as the type of devices and the OS framework under evaluation did not exist before 2007. No studies were excluded based on their language of publication.

2.4.4. Data collection and analysis

2.4.4.1. Selection of studies

An initial search included studies published up to 02 July 2014, and an update search included studies published up to 13 April 2015 (Appendices 15 and 16). The citations identified through this process were imported into the reference management software EndNote, and duplicate records of the same report were removed using the built-in function offered by this software.

Overall, the title and abstract of 22,676 citations were independently examined by two reviewers to identify potentially relevant studies. Subsequently, the full-text reports of 264 citations were independently assessed by two independent reviewers for compliance with the review's inclusion and exclusion criteria.

The two independent reviewers met to reconcile their independent assessments, and to produce a final list of included studies. Any disagreements between the two reviewers was resolved through discussion, without the need for arbitration from a third independent reviewer.

Excluded studies and the reason for exclusion were summarised in a Characteristics of Excluded Studies table (Appendix 17).

2.4.4.2. Data extraction and management

Data from the included studies were extracted by two independent reviewers using a structured web-based form. Where available, the following information was extracted from each record: general study details; study methods; number and description of participants; type of survey used, and their delivery modes; outcomes measured (including time points); study findings; and study conclusions.

The two independent reviewers met to compare and unify their structured web-forms. Any disagreements between the two reviewers was resolved through discussion, without the need for arbitration from a third independent reviewer. The final data extraction form was presented as Characteristics of Included Studies tables (Appendices 18 to 31).

2.4.4.3. Risk of bias assessment

Risk of bias in RCTs was assessed using the Cochrane's tool for assessing the risk of bias in randomised trials [21]. This tool assesses the presence of bias across the following constructs: random sequence generation; allocation concealment; blinding of participants and personnel; blinding of outcome assessment; incomplete outcome data; selective outcome reporting; and other bias (including imbalance of outcome measures at baseline, comparability between the characteristics of the intervention and control group, and protection against contamination).

As recommended by the Cochrane Handbook [21], the risk of bias in crossover trials was assessed across the following constructs: suitability of the crossover design; evidence of a carry-over effect; whether only first period data were available; incorrect statistical analysis; and comparability of results with those from non-randomised trials.

No cluster RCTs were identified that met the inclusion criteria for this systematic review.

The risk of bias assessment was conducted as part of the data extraction process, and the relevant tools were embedded within the structured web-form. Each domain was classified as having Low, High or Unclear risk of bias. The results of this assessment are summarised in the Characteristics of Included Studies tables (Appendices 18 to 31).

2.4.4.4. Measures of effect

The effect of continuous outcomes (i.e., comparison of mean scores between delivery modes, data completeness, time taken to complete a survey, and acceptability) was assessed in relation to the mean difference (MD) and 95% confidence intervals (CI).

The effect of dichotomous outcomes (i.e., acceptability) was assessed in relation to the odds ratio (OR) and 95% CI.

2.4.4.5. Missing data

The authors of studies with missing or incomplete data were contacted to request the missing information. However, no replies were received.

2.4.4.6. Assessment of heterogeneity

Clinical and methodological diversity between the included studies across all systematic review outcomes was assessed qualitatively. Clinical diversity was assessed in relation to the type of device and platform used to deliver surveys, the intensity of the sampling protocols, and the characteristics of the participants. Methodological diversity was assessed in relation to the properties of a survey, study methodology, and definition of outcomes.

As part of this assessment, it was concluded that a meta-analysis would be appropriate only for studies that analysed participants falling within the same age group, used validated surveys, and chose the same comparators. However, only a small number of included studies met these criteria. In addition, included studies still displayed substantial clinical and methodological diversity, even after taking into consideration these criteria. Therefore, a meta-analysis was not performed and no formal statistical test was used to quantify statistical heterogeneity.

2.4.4.7. Reporting biases

To minimise the risk of publication bias, a comprehensive search of multiple bibliographic databases, trials registries and grey literature repositories, reference lists of potentially relevant studies, and the public profiles of authors with relevant publications was conducted. Since there were fewer than 10 studies in each of the analyses, the assessment of reporting bias using a funnel plot regression weighted by the inverse of the pool variance was not required.

2.4.4.8. Data synthesis

Since performing a meta-analysis was not indicated, the results were analysed using a narrative synthesis approach. The framework proposed by Rodgers and colleagues [22] was adapted to guide this process.

The initial data synthesis plan included using the Grading Recommendations Assessment, Development and Evaluation (GRADE) to assess the quality of the evidence, and to determine the magnitude of the effect size and the sum of the available data on the systematic review

outcomes to produce a Summary of Findings table. However, this was not implemented as a meta-analysis was not conducted.

The original systematic review protocol outlined a complete methodological plan to cover many scenarios. However, in conducting the systematic review, many of the steps mentioned in this protocol became irrelevant: risk of bias assessment and unit of analysis issues in cluster RCTs (as no cluster RCTs that met the review's inclusion criteria were identified); summary of findings table (as a meta-analysis was not performed); subgroup analysis (as a meta-analysis was not performed) and sensitivity analysis (as none of the conditions outlined in the systematic review protocol were met).

2.5. Results

2.5.1. Description of studies

2.5.1.1. Results of the search

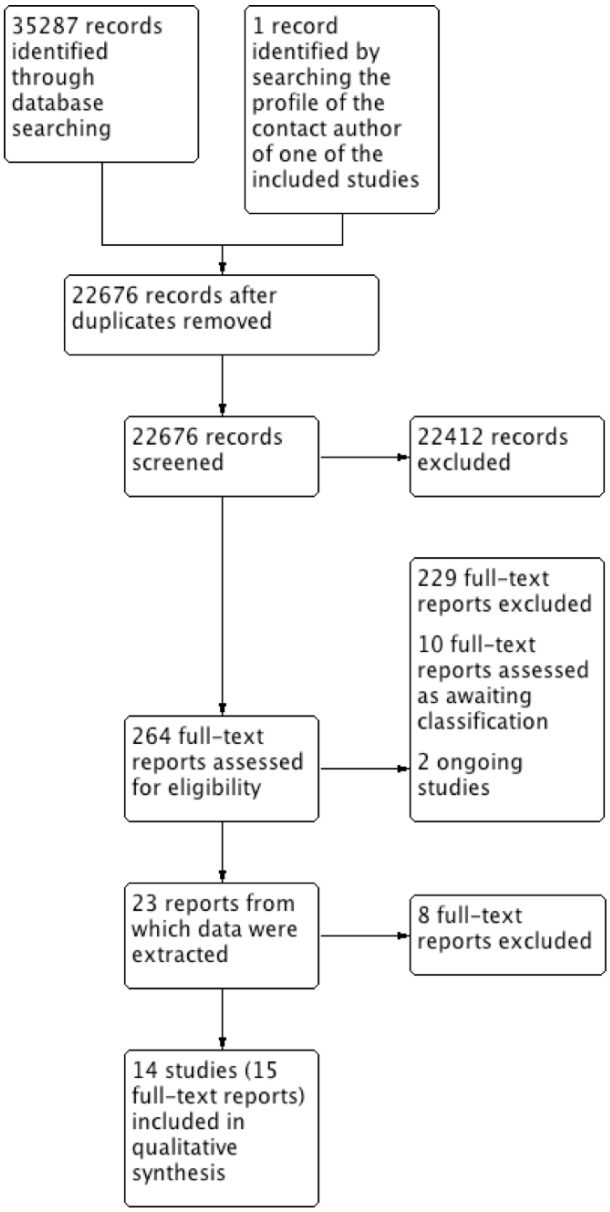
Overall, the titles and abstracts of 22,676 records were screened. Of these, 22,412 records were excluded, and the full-text reports of 264 were retrieved and assessed for eligibility. Of these, 229 records were excluded, 10 were categorised as Awaiting classification (as there was insufficient information available from the full-text reports, and any attempts to contact the authors were unsuccessful), two corresponded to ongoing trials, and eight were excluded (as the data extraction process revealed that they did not meet the inclusion criteria for this systematic review).

Overall 15 records were included, of which nine were papers published in peer-reviewed journals and six were conference posters or abstracts. Since the unit of analysis in a systematic review is the study and not the reports, further adjustments were made. Two records corresponded to a single study and were included as such. Two other records also corresponded to a single study; however, they conducted two separate analysis on two different samples of participants. Therefore, these two records were included as two studies, with both records contributing data to each study.

Therefore, 14 studies were included in this systematic review: Ainsworth and collegues [23]; Brunger and colleagues [24]; Bush and colleagues [25]; Garcia-Palacios and colleagues [26];

Khraishi and colleagues [27, 28]; Kim and colleagues [29]; Lamber and colleagues [30]; Newell and colleagues [31]; Salaffi and colleagues [32]; Schemmann and colleagues [33]; Sigaud and colleagues [34]; Stomberg and colleagues [35]; Sun and colleagues (which accounted for two studies) [36, 37]. These studies assessed 2,275 participants (although only 2,272 participants were included in the statistical analyses as there was missing data for three of them). See Figure 1 for additional information.

Figure 1. PRISMA flow diagram



2.5.1.2. Characteristics of included studies

This section presents a description of the included studies according to the following domains: types of studies; types of data; types of method; types of comparisons; and types of outcome measures.

2.5.1.2.1. Types of study

All the studies were conducted in high income countries: Canada [27, 28, 36, 37]; France [34]; Germany [33]; Italy [30, 32]; Republic of Korea [29]; Spain [26]; Sweden [35]; UK [23, 24]; and the United States (US) [25, 31].

Three studies conducted RCTs [30, 31, 35], ten conducted crossover trials [23, 25-29, 32-34, 36, 37] and one conducted a paired repeated measures study [24]. The duration of these studies varied between one day [24, 30-33, 36, 37] and six months [34]. Washout periods in crossover trials varied between 30 minutes [36, 37] and one week [23, 29].

The main aim of the included studies were: to compare the psychometric properties of a survey between different delivery modes [23-28, 31-33, 36, 37]; to develop a smartphone application for delivering a validated survey (demonstrating its psychometric properties) [29]; to demonstrate the data equivalence between different delivery modes whilst assessing the impact that patient-related characteristics can have on usability [30]; and to evaluate the performance of a new delivery mode for recording patient data [34, 35].

Three studies provided incentives to their participants in the form of phone credit [23], three weeks of free psychological treatment [26], and a gift card [31].

2.5.1.2.2. Types of data

The data gathered in the included studies were classified across the following areas: properties of the self-administered surveys; characteristics of the sample of participants; and setting in which survey completion took place.

2.5.1.2.2.1. Properties of the self-administered survey

For each survey, the following properties were considered: validation status; clinical uses; and type of response scale.

In relation to their validation status and clinical application, nine studies used validated surveys [24, 27-33, 36, 37]. Brunger and colleagues [24] developed a visual analogue scale (VAS) to evaluate satiety before and after the consumption of a high- or a low-energy drink. Khraishi and colleagues [27, 28] used the Health Assessment Questionnaire (HAQ) developed by Bruce and Fries [38], which assesses functional status in rheumatological conditions across five patient-related dimensions: avoiding disability; being free of pain or discomfort; avoiding adverse treatment effects; keeping treatment costs low; and postponing death. Kim and colleagues [29] used the International Prostate Symptom Score (IPSS) developed by Barry and colleagues [39, 40] to assess symptoms of benign prostatic hyperplasia: sensation of incomplete bladder emptying, urinary frequency, urinary intermittency, difficulty urinating, strength of the urinary stream, straining, nocturia, and quality of life. Lamber and colleagues [30] administered the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire C-30 (EORTC QLQ – 30) [41] to assess quality of life in cancer patients. Newell and colleagues [31] administered the Center for Epidemiologic Studies Depression Scale (CES-D) [42, 43] and the Regulatory Focus Questionnaire (RFQ) [44]. The former measures dysphoria, anhedonia, appetite, sleep, thinking, feelings of worthlessness, fatigue, agitation, and suicidal ideation in patients with suspected depression. The RFQ measures an individual's orientation towards her/his goal across two subscales: prevention (focusing on safety and responsibility) and promotion (focusing on hopes and accomplishments). Salaffi and colleagues [32] used the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) [45] and the Bath Ankylosing Spondylitis Functional Index (BASFI) [46]. The former is a 10-centimetre VAS that measures the severity of fatigue, spinal and peripheral joint pain, localised tenderness and morning stiffness in patients with ankylosing spondylitis. The BASFI is a two-part survey that measures function and ability to perform activities of daily living (ADL) in this group of patients. Schemmann and colleagues [33] administered the German version of the short International Hip Outcome Tool (iHOT-12) developed by Griffin and colleagues [47] to assess the quality of life and functional status in patients with hip disorders.

Sun and colleagues [36, 37] used the Faces Pain Scale Revisited (FPS-R) and the Color Analog Scale (CAS) to assess post-surgical pain in paediatric patients.

Two studies developed composite scales, which were derived from already validated instruments [23, 25]. Ainsworth and colleagues [23] developed a diagnostic tool from a mobile phone assessment scale that assesses symptoms of hopelessness, depression, hallucinations, anxiety, grandiosity, paranoia and delusion. This scale was administered through EMA techniques. Bush and colleagues [25] developed a scale called Mobile Screener, which incorporates the Post Traumatic Stress Disorder Checklist (PTSD-Checklist), PHQ-9, Revised Suicidal Ideation Scale (R-SIS), Deployment Risk and Resilience Inventory-Unit Support (DRRI-US), Dimensions of Anger 5 (DAR5), Sleep Evaluation Scale, and TBI Self Report of Symptoms.

Three studies used non-validated scales [26, 34, 35]. Sigaud and colleagues [34] used a non-validated diary to monitor treatment compliance in patients with Haemophilia A. Stomberg and colleagues [35] monitored post-surgical pain using a non-validated, numerical rating scale (NRS). Lastly, Garcia-Palacios and colleagues [26] used EMA techniques to collect patient reported outcome measures (PROMs) of pain, fatigue and mood through an NRS.

In relation to the response scales used, Bush and colleagues [25] used categorical response scale: their Sleep Evaluation Scale required true or false answers to ten items. The scales used in the remaining studies consisted of continuous scales (including VAS) [24, 32, 36, 37], NRS [26, 31, 32, 35], adjectival or Likert scales [23, 25, 29, 30, 33], and face scales [36, 37].

2.5.1.2.2.2. Characteristics of the sample of participants

Participants recruited into the included studies were assessed according to their health status and age group.

Most included studies (11 out of 14) recruited participants from clinical populations: rheumatology [26-28, 32, 33]; surgery [35-37]; psychiatry [23]; urology [29]; oncology [30]; and haematology [34]. The remaining three studies recruited healthy volunteers [24, 25, 31].

Concerning participants' age group, only 2 studies recruited participants from a paediatric population: children aged between four and 11 years old, and between five and 18 years old [36, 37]. The remaining 12 studies recruited adult participants aged between 18 and 80 years old.

2.5.1.2.2.3. Setting of survey completion

Four studies asked participants to complete surveys in a naturalistic setting [23, 26, 34, 35]. These studies also implemented more frequent and longer sampling protocols. The remaining ten studies administered surveys in clinical settings [24, 25, 27-33, 36, 37].

2.5.1.2.3. Types of method

This section refers to the use of apps as a delivery mode for self-administered surveys, which were assessed according to the following characteristics: devices and technical platforms used; the main functionality offered by the apps; factors related to human-computer interaction; survey administration protocols; and additional interventions offered as part of the study.

2.5.1.2.3.1. Devices and technical platforms

Lambert and colleagues [30] asked their participants to complete the EORTC QLQ-30 using apps running on both smartphones and tablet computers. While the model of the tablet computers used was not specified, the model of the smartphones was Nokia N97 running Symbian OS. Their app, MobiDay, was developed specifically for smartphones, which suggests that the tablet computers might not be compatible with this review's inclusion criteria (therefore, data related to the use of their tablet computer was considered as a control group).

Eight studies used apps running on smartphones to administer their surveys. The app used in Bush and colleagues [25] ran on iOS devices, while the app in Stomberg and colleagues [35] supported iOS, Android and Java-enabled devices. Ainsworth and colleagues [23] used Orange San Francisco handsets running Android OS, and Garcia-Palacios and colleagues [26] used HTC 1 Diamond handsets running Windows Mobile OS. Kim and colleagues [29], Sigaud and

colleagues [34], and Sun and colleagues [36, 37] reported having used smartphones in their studies; however, they did not specify the model of the handsets.

The remaining five studies used apps running on tablet computers. Both Khraishi and colleagues [27, 28] and Newell and colleagues [31] used iPads® (iOS), while Brunger and colleagues [24] used iPad minis™ (iOS). On the other hand, Salaffi and colleagues [32] used an Archos 101 tablet computer running Android OS. Although Schemmann and colleagues [33] reported having used tablet computers in their study, they did not specify the model.

2.5.1.2.3.2. Main functionality

Only four studies provided a list of the main functionality offered by their apps. Ainsworth and colleagues [23] allowed the configuration of the number of questions that were presented to participants on each day, as well as the times at which these were displayed. They also permitted the configuration of questions, creation of question sets, question branching, questionnaire timeout, time stamping of responses, and skip procedures (i.e., dependency between questions). Garcia-Palacios and colleagues [26] allowed the configuration of the number of questions displayed on each day, and time stamping of responses. Stomberg and colleagues [35] implemented configurable questions and allowed the configuration of the number of questions that were presented on each day. Salaffi and colleagues [32] implemented skip procedures and compulsory questions.

2.5.1.2.3.3. Human-computer interaction

Seven studies reported on the human-computer interaction factors that were implemented in their apps. Ainsworth and colleagues [23] allowed their participants to set their own schedule of alerts. These alerts were delivered at semi-random time intervals, and participants were able to snooze them for five minutes. Only one question was displayed on the screen at any given time; however, participants were able to navigate back and forth between questions. Responses were captured through a continuous slider bar mapped onto a seven-point Likert scale, and were saved as soon as they were entered. Brunger and colleagues [24] presented their participants with one question per screen at any given time. Responses were captured through a continuous slider bar mapped onto a 10-cm horizontal

line on a touchscreen display. Responses were automatically transmitted to a secure database via a wireless connection. Garcia-Palacios and colleagues [26] implemented alerts in the form of audio signals. Audio reminders were played if participants did not enter an answer within the time specified by the initial alert. Thereafter, reminders were played every minute during the first 15 minutes, and every 15 minutes for the next hour. This app also featured audio-recorded instructions. Stomberg and colleagues [35] implemented alerts in the form of push notifications. These alerts were delivered every four hours, and SMS reminders were sent to participants if they had not entered a response within 13 minutes of an alert.

Kim and colleagues [29], Lamber and colleagues [30], Salaffi and colleagues [32], and Stomberg and colleagues [35] presented one question per screen at any given time. However, only Kim and colleagues [29] allowed their participants to navigate back and forth between questions, and to modify their answers before submitting them. Participants in this study were required to confirm the selected response option before it was saved. On the other hand, responses were saved automatically in Lamber and colleagues [30], Salaffi and colleagues [32], and Stomberg and colleagues [35]. Moreover, participants in Lamber and colleagues [30] were able to save their partially completed surveys, and resume survey completion at a later time. Lastly, Salaffi and colleagues [32] implemented voice and text synchronisation, and replay buttons for each question stem and response option.

2.5.1.2.3.4. Survey administration protocol

Participation in ten studies was limited to one day [24, 25, 27-33, 36, 37]. The remaining four studies implemented longer sampling protocols. Garcia-Palacios and colleagues [26] asked participants to complete surveys three times a day for seven days. Both Ainsworth and colleagues [23], and Stomberg and colleagues [35] asked participants to complete surveys four times a day for six days. Lastly, Sigaud and colleagues [34] asked their participants to keep a diary for three months, although the frequency of entries required was not specified in the study report.

2.5.1.2.3.5. Additional interventions

Only four studies reported implementing additional interventions. Ainsworth and colleagues [23], Newell and colleagues [31], Salaffi and colleagues [32], and Stomberg and colleagues [35] offered training on the use of their app or the device. Both Ainsworth and colleagues [23] and Stomberg and colleagues [35] had telephone contact with their participants during the sampling period. Moreover, staff members in Stomberg and colleagues [35] were in charge of the app installation. Lastly, Ainsworth and colleagues [23] conducted the Positive and Negative Syndrome Scale (PANSS) semi-structured interview, both before and after the sampling period.

2.5.1.2.4. Types of comparisons

The comparators evaluated in the included studies were assessed along the same dimensions as apps running on smartphones or tablet computers.

2.5.1.2.4.1. Devices and technical platforms

Bush and colleagues [25], and Lamber and colleagues [30] used both laptop computers (a MacBook Pro in the case of the latter) and paper surveys as their comparators. In addition, the tablet computer used in Lamber and colleagues [30] was considered a comparator, as it is unlike to match this systematic review's criteria for the type of method. Brunger and colleagues [24] used an iPAQ PDA as their comparator, and Ainsworth and colleagues [23] used SMS-based surveys. Sun and colleagues [36, 37] compared their app to a version of the CAS that was printed on a plastic ruler. The remaining included studies chose paper as their comparator [26-29, 31-37].

2.5.1.2.4.2. Main functionality

Ainsworth and colleagues [23] permitted the configuration of the number of questions that were sent to participants on each day. They also implemented configurable questions, multiple question sets, question branching, questionnaire timeout, time stamping of responses, and skip sequences.

The remaining included studies using a technology-based comparator did not report on the main functionality offered.

2.5.1.2.4.3. Human-computer interaction

SMS alerts in Ainsworth and colleagues [23] were delivered at semi-random intervals (one per question). A SMS reminder was sent if no response was received within five minutes of the initial alert. Each SMS corresponded to only one question, and answers were submitted by responding to the SMS alert (or reminder). Subsequent questions were sent only when an answer to the current question had been submitted. Brunger and colleagues [24] presented one question per screen at any given time, and responses were captured through a continuous slider mapped onto a 64-mm horizontal line. Answers had to be entered using a stylus. Participants allocated to the tablet computer in Lamber and colleagues [30] were also asked to enter their response with a stylus. Those allocated to the laptop computer had to enter their answers through an application via a web browser. This application adapted its graphical interface to the characteristic of the device, and presented one question per screen at any given time. Responses could be selected from a drop-down list.

2.5.1.2.4.4. Survey administration protocol

Except for Stomberg and colleagues [35] (who sampled participants in the comparison group four times a day for four days), the survey administration protocols between apps and their comparators were identical.

2.5.1.2.4.5. Additional interventions

Ainsworth and colleagues [23] offered training, and had telephone communication with their participants during the sampling period. They also conducted the PANSS semi-structured interview both before and after the sampling period. Newell and colleagues [31] offered training on the use of an iPad® to all participants (including those allocated to the paper version of the survey). Salaffi and colleagues [32] offered on-site assistance to their participants.

2.5.1.2.5. Types of outcome measures

2.5.1.2.5.1. Data equivalence

Most studies (11 out of 14) measured data equivalence. Bush and colleagues [25], Kim and colleagues [29], and Salaffi and colleagues [32] compared mean scores between delivery modes, and calculated ICC coefficients. Ainsworth and colleagues [23], Garcia-Palacios and colleagues [26], Khraishi and colleagues [27, 28], and Newell and colleagues [31] compared mean scores between delivery modes only. Sun and colleagues [36, 37] compared mean scores and calculated the Pearson correlation coefficient. Schemmann and colleagues [33] calculated the ICC coefficient. Brunger and colleagues [24] calculated correlation coefficients for each of five question; however, the type of coefficient was not specified in the study report.

2.5.1.2.5.2. Data accuracy

None of the included studies measured data accuracy.

2.5.1.2.5.3. Data completeness

Ainsworth and colleagues [23], and Garcia-Palacios and colleagues [26] measured data completeness. The former measured it as the mean number of daily entries. Garcia-Palacios and colleagues [26] defined this outcome as the difference in the mean number of complete and incomplete records.

2.5.1.2.5.4. Response rates

None of the included studies measured response rates.

2.5.1.2.5.5. Time needed to complete a survey

Ainsworth and colleagues [23], Khraishi and colleagues [27, 28], and Salaffi and colleagues [32] defined this outcome as the difference in the mean time taken to complete a survey between an app and the corresponding comparators.

2.5.1.2.5.6. Adherence to a survey administration protocol

Ainsworth and colleagues [23], Sigaud and colleagues [34], and Stomberg and colleagues [35] measured adherence to a pre-specified survey administration protocol. Ainsworth and colleagues [23] defined it as the difference between two delivery modes in the proportion of individuals who completed at least a third of all possible data points. Sigaud and colleagues [34] defined it as the proportion of individuals responding to the data entry alerts. Lastly, Stomberg and colleagues [35] defined this outcome as the proportion of participants following the pre-specified data collection protocol.

2.5.1.2.5.7. Acceptability to respondents

Most studies (12 out of 14) measured the acceptability to respondents of the delivery mode. However, each study used their own bespoke measure. Therefore, there was considerable variability across the included studies in terms of the definition of acceptability: preference [23, 25-29, 31-33, 36, 37]; ease of use [23, 26-28, 30, 31, 33]; willingness to use a delivery mode [29, 34]; satisfaction [30, 31, 34]; informativeness of the system [26, 30, 31]; perceived time taken to complete surveys [26-28]; perceived benefit [27, 28]; perceived usefulness [26]; perceived ability to complete a survey using a given delivery mode [31]; maximum duration that participants would be willing to use a delivery mode [23]; and reactivity to a delivery mode and successful integration into participants' routines [23].

2.5.2. Risk of bias

2.5.2.1. Randomised controlled trials

The risk of bias in Lamber and colleagues [30], Newell and colleagues [31], and Stomberg and colleagues [35] was assessed using the Cochrane Collaboration's Risk of Bias assessment tool for RCTs [21]. A summary of this assessment is presented in Figures 2 and 3.

Figure 2. Risk of bias graph for RCTs

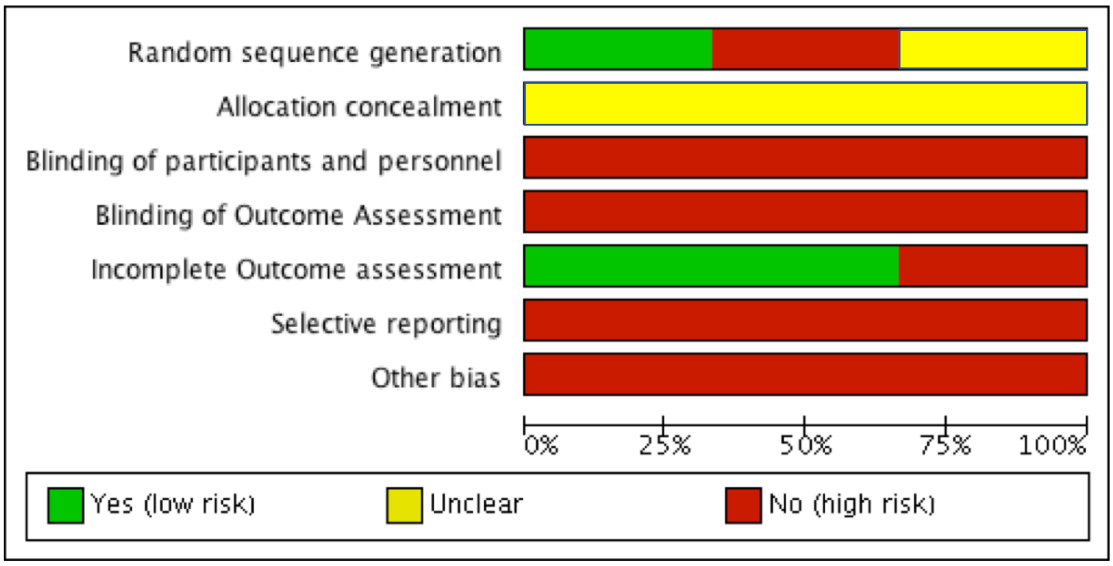
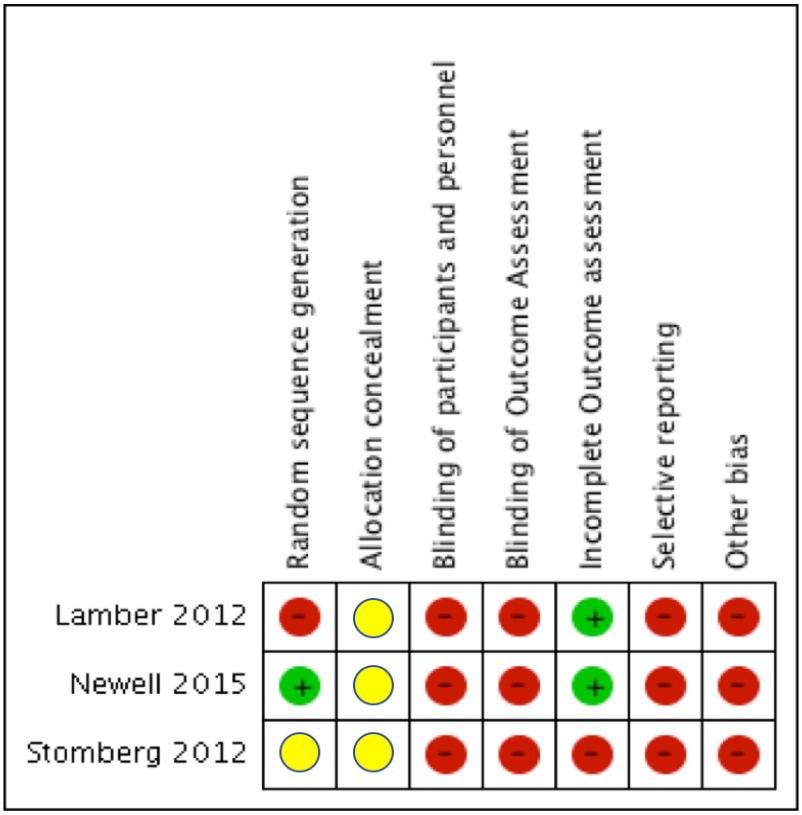


Figure 3. Risk of bias summary for RCTs



2.5.2.1.1. Selection bias

2.5.2.1.1.1. Random sequence generation

Participants in Lamber and colleagues [30] were selected by clinicians, and were randomly allocated to one of four experimental conditions. However, the method by which this was done was not specified in the study report. However, 47.3% of participants were allocated to the laptop group alone. For this reason, its risk of selection bias due to random sequence generation was considered high.

Newell and colleagues [31] conducted computerised randomisation using Qualtrics software. Therefore, its risk of bias on this domain was assessed as low.

Lastly, participants in Stomberg and colleagues [35] were randomly allocated to one of two experimental groups; however, the technique used to achieve this was not specified in the study report. For this reason, its risk of bias on this domain was considered unclear.

2.5.2.1.1.2. Allocation concealment

There was insufficient information in the study report of the three studies that conducted an RCT to assess their risk of selection bias due to allocation concealment. For this reason, this domain was judged as unclear.

2.5.2.1.2. Blinding

2.5.2.1.2.1. Blinding of participants and personnel (performance bias)

Blinding of participants and personnel was not possible in these circumstances, as the delivery mode could not be concealed. Awareness of the delivery modes being offered to other participants could have influenced participants' motivation and performance during survey completion. Therefore, the risk of performance bias for these three studies was considered high.

2.5.2.1.2.2. Blinding of outcome assessment (detection bias)

The reports of the three studies that conducted an RCT did not specify if outcome assessors were blinded to participant allocation. However, manual data entry and manual calculation

of scores for paper surveys could have introduced detection bias. For this reason, the risk of bias on this domain for these studies was considered high.

2.5.2.1.3. Attrition bias

All the participants in Lamber and colleagues [30], and Newell and colleagues [31] completed the study procedures, and their data were included in the final statistical analyses. Therefore, their risk of attrition bias due to incomplete or missing data was low.

Stomberg and colleagues [35] did not include data from three participants in their analysis. For this reason, its risk of attrition bias due to incomplete or missing data was considered high.

2.5.2.1.4. Reporting bias

Lamber and colleagues [30] evaluated the impact that a patient's profile had on the usability of electronic delivery modes. However, they only focused on those participants allocated to the laptop computer group. Participants in Newell and colleagues [31] underwent randomisation before completing each of two surveys. In between these surveys, participants were asked to complete a measure of acceptability. However, in the statistical analysis of this outcome, the authors included data from participants who completed both surveys on the same delivery mode. Moreover, only participants recruited from the second community where the study took place were asked to complete the BRIEF health literacy scale and other items concerning the survey format. Lastly, Stomberg and colleagues [35] measured response rates. However, this outcome was defined as compliance with the original sampling protocol, and was reported on a day-by-day basis. In addition, the authors attempted to report a comparison of overall pain scores, sometimes according to type of surgery and sometimes according to the delivery mode. However, the lack of appropriate reporting makes it difficult to identify the meaningful comparisons. For these reasons, their risk of reporting bias was considered high.

2.5.2.1.5. Other sources of bias

Lamber and colleagues [30] did not report the standard deviation (SD) of their usability scores. Participants in Newell and colleagues [31] received training on how to use an iPad®, regardless of their allocation. Participants allocated to the app group in Stomberg and colleagues [35] received training on pain management and on the use of the app. Moreover, the sampling period between the two experimental groups in this study was different. For these reasons, their risk of other bias was judged as high.

2.5.2.2. Crossover studies

The risk of bias in the remaining 11 included studies was assessed according to the recommendations outlined in the Cochrane Handbook for Systematic Reviews of Interventions [21] for this type of study design. A summary of this assessment is presented in Figures 4 and 5.

Figure 4. Risk of bias graph for crossover trials

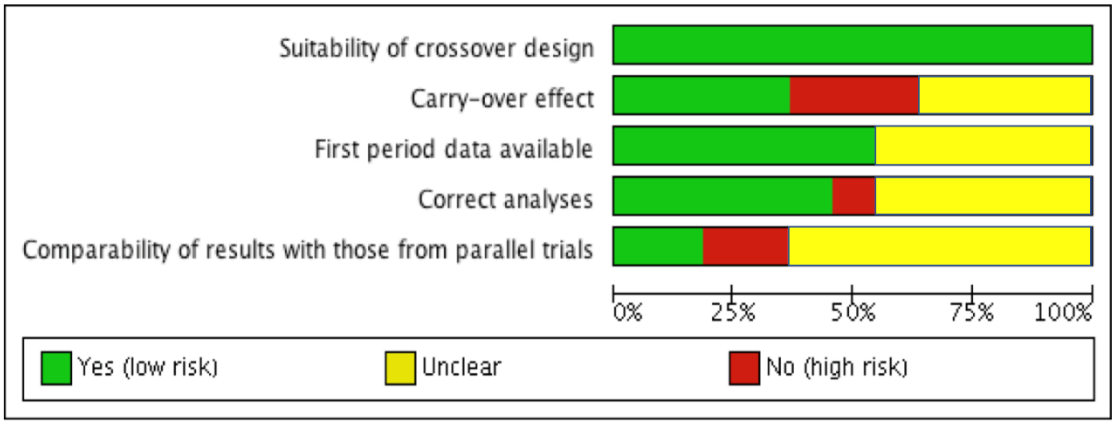


Figure 5. Risk of bias summary for crossover trials



2.5.2.2.1. Suitability of the study design

Crossover trials are recognised as a suitable methodology for the assessment of data equivalence between delivery modes [9, 10]. Therefore, the risk of bias on this domain for all 11 studies was considered low.

2.5.2.2.2. Carry-over effect

Ainsworth and colleagues [23] assessed the interaction between the sampling period and method of assessment; they found that the order of the two experimental conditions did not significantly predict the number of entries made, or the time need to complete each entry. Similarly, Sigaud and colleagues [34] assessed the interaction between the sampling period and method of assessment. They found that the sequence of the two diaries, and the specific effect related to the patient, did not influence the rate of diary completion. Although Kim and colleagues [29] did not formally test the presence of a carry-over effect, they implemented a one-week washout period to minimise its likelihood. Brunger and colleagues [24] counterbalanced the order in which delivery modes were used to administer their surveys, and accounted for the type of device in their data analysis. Therefore, their risk of bias on this domain was assessed as low.

Bush and colleagues [25], Garcia-Palacios and colleagues [26], and Salaffi and colleagues [32] did not test for the occurrence of carry-over effect, and implemented very short washout periods. For these reasons, their risk of bias on this domain was judged as high.

The risk of bias on this domain was judged as unclear for Khraishi and colleagues [27, 28], Schemmann and colleagues [33], and Sun and colleagues [36, 37], as there was insufficient information available in their study reports.

2.5.2.2.3. Only first period data available

Ainsworth and colleagues [23], Brunger and colleagues [24], Bush and colleagues [25], Garcia-Palacios and colleagues [26], Kim and colleagues [29], and Salaffi and colleagues [32] included data from both sampling periods in their analyses. Therefore, their risk of bias on this domain was considered low.

The risk of bias on this domain was judged as unclear for Khraishi and colleagues [27, 28], Schemmann and colleagues [33], Sigaud and colleagues [34] and Sun and colleagues [36, 37], as there was insufficient information available in their study reports.

2.5.2.2.4. Suitability of the statistical analysis

Ainsworth and colleagues [23], Brunger and colleagues [24], Bush and colleagues [25], Kim and colleagues [29], and Salaffi and colleagues [32] used measures (i.e., comparison of mean scores or a wide range of correlation coefficients) and statistical techniques that are suitable for the assessment of data equivalence between alternative delivery modes [9, 10]. For this reason, their risk of bias on this domain was considered low.

Garcia-Palacios and colleagues [26] implemented appropriate statistical techniques. However, they did not report data on mood assessment, and the data from seven participants were excluded from the final analysis. Therefore, its risk of bias on this domain was judged as high.

The risk of bias on this domain was judged as unclear for Khraishi and colleagues [27, 28], Schemmann and colleagues [33], Sigaud and colleagues [34] and Sun and colleagues [36, 37], as there was insufficient information available in their study reports.

2.5.2.2.5. Comparability of results with those from parallel trials

Ainsworth and colleagues [23] followed appropriate randomisation procedures, tested for the presence of carry-over effects, and implemented a one-week washout period. Therefore, their results are comparable with those from parallel trials (i.e., low risk of bias on this domain).

Brunger and colleagues [24] recruited a small sample of participants. Bush and colleagues [25] did not test for the presence of carry-over effects, and implemented a 90-minute washout period. For these reasons, the results from these studies might not be comparable to those from parallel trials (i.e., high risk of bias on this domain).

The washout period in Salaffi and colleagues [32] was only 60 minutes. Although they claimed having implemented measures to minimise the likelihood of carry-over effects, they not

specified them in their study report. Therefore, its risk of bias on this domain was assessed as unclear.

The risk of bias on this domain was judged as unclear for Garcia-Palacios and colleagues [26], Khraishi and colleagues [27, 28], Kim and colleagues [29], Schemmann and colleagues [33], Sigaud and colleagues [34] and Sun and colleagues [36, 37], as there was insufficient information available in their study reports.

2.5.3. Effect of delivery mode

This systematic review assessed five out of seven outcomes, as none of the included studies measured data accuracy or response rates. The results concerning data equivalence, data completeness, time need to complete a survey, adherence to survey administration protocols, and acceptability to respondents were reported separately. Each of these outcomes were reported according to the setting in which survey completion took place: clinical settings, and naturalistic settings. Within each setting, outcomes were reported according to the type of delivery modes being compared: app versus paper, app versus laptop computer, app versus tablet computer (pre-2007), app versus PDA, app versus a plastic ruler, and app versus SMS.

A qualitative assessment of heterogeneity highlighted considerable levels of clinical and methodological diversity between the included studies. Regarding the former, nine studies used smartphones and five used tablet computers. In addition, five studies developed apps for iOS, three studies for Android OS, one study for Windows Mobile OS, and one study for Symbian OS. Differences in the technical specifications of these devices, and in the user interfaces provided by different OS, could affect users' interaction with the survey and thus the response generation process [1, 4, 15]. Included studies also varied in terms of the survey administration protocols they implemented. Ten studies sampled participants for one day, while three studies sampled participants (three or four times a day) for a week and one study for three months. Concerning the characteristics of participants, only two studies assessed a paediatric sample of participants (between four and 18 years old). The remaining 12 studies recruited adult participants. Three studies recruited healthy volunteers, while the remaining

11 studies recruited participants from a wide range of clinical populations: psychiatry, rheumatology, surgery, urology, oncology, and haematology.

Concerning methodological diversity, nine studies implemented validated scales, two studies developed and administered composited scales derived from validated instruments, two studies used non-validated scales, and one study used PROMs with unclear validation status. Validated instruments differed in their clinical applications: functional status assessment, pain assessment, mental health status assessment, symptom scores, assessment of individual differences, food consumption and satiety, and health-related quality of life. Moreover, ten studies used a crossover study design, three studies were RCTs, and one study conducted a paired repeated-measures crossover study design. The delivery modes against which apps were compared included (in order of prevalence) paper, laptop computers, SMS, PDA, and a plastic ruler. Four studies were conducted in naturalistic settings, while the remaining ten were conducted in clinical settings. Lastly, outcome definitions varied considerable across the included studies, with acceptability to respondents displaying the highest degree of variability.

Furthermore, the risk of bias assessment highlighted some concerns regarding the methodological quality of the included studies. Two of the three RCTs were found to have high risk of bias in five out of seven domains. The remaining RCT had high risk of bias in four domains. In addition, there was considerable uncertainty (i.e., unclear risk of bias) in most of the crossover trials.

For these reasons, a meta-analysis was not performed. Instead, the results were interpreted using a narrative synthesis approach.

2.5.3.1. Clinical settings

2.5.3.1.1. App versus paper

2.5.3.1.1.1. Data equivalence

Concerning the administration of validated surveys in an adult population, Bush and colleagues [25] did not observe any difference between an app and paper in the mean responses given to seven symptom dimensions:

- $MD_{PCL\text{-}C}$ = 2.60, 95% CI -3.17 to 8.37
- $MD_{PHQ\text{-}9}$ = -0.70, 95% CI -2.93 to 1.53
- $MD_{RSI\text{-}S}$ = 0.30, 95% CI -0.24 to 0.84
- $MD_{DRRI\text{-}US}$ = 1.90, 95% CI -4.17 to 7.97
- MD_{Anger} = 1.40, 95% CI -0.52 to 3.32
- MD_{Sleep} = -0.10, 95% CI -1.26 to 1.06
- MD_{TBI} = 0.40, 95% CI -0.43 to 1.23

These authors also found that, except for anger, the ICC coefficient for all symptom dimensions exceeded 0.70:

- $ICC_{PCL\text{-}C}$ = 0.90, 95% CI 0.85 to 0.96
- $ICC_{PHQ\text{-}9}$ = 0.92, 95% CI 0.87 to 0.96
- $ICC_{RSI\text{-}S}$ = 0.86, 95% CI 0.79 to 0.94
- $ICC_{DRRI\text{-}US}$ = 0.81, 95% CI 0.71 to 0.91
- ICC_{Anger} = 0.67, 95% CI 0.51 to 0.93
- ICC_{Sleep} = 0.95, 95% CI 0.92 to 0.98
- ICC_{TBI} = 0.88, 95% CI 0.82 to 0.95

Khraishi and colleagues [27, 28] found no difference between their app and a paper version of the HAQ: 95% CI -0.159 to 0.345, P = 0.459.

Kim and colleagues [29] observed no difference between an app and paper in the overall IPSS and IPSS Quality of Life scores:

- MD_{IPSS} = -0.01, 95% CI -0.55 to 0.53
- $MD_{IPSS\ QoL}$ = 0.00, 95% CI -0.10 to 0.10

These authors also found that the ICC coefficient of the overall IPSS score exceeded 0.70: 0.935, 95% CI 0.927 to 0.941, P < 0.001.

Newell and colleagues [31] found no difference between an app and paper in the mean scores of three surveys:

- CES-D: $Mean_{App}$ = 1.21 (SD = 0.54); $Mean_{Paper}$ = 1.10 (SD = 0.54); t = 1.33; P = 0.19

- RFQ-Promotion: $Mean_{App}$ = 19.85 (SD = 3.04); $Mean_{Paper}$ = 20.09 (SD = 3.04); t = 0.53; P = 0.59
- RFQ-Prevention: $Mean_{App}$ = 15.93 (SD = 3.37); $Mean_{Paper}$ = 16.19 (SD = 3.37); t = 0.48; P = 0.63

Salaffi and colleagues [32] found no difference between and app and paper in the mean BASFI and BASDAI scores:

- MD_{BASFI} = -0.01, 95% CI -0.91 to 0.89
- MD_{BASDAI} = 0.05, 95% CI -0.80 to 0.90

These authors also found that the ICC coefficient for each of these instruments exceeded 0.70:

- ICC_{BASFI} = 0.90, 95% CI 0.88 to 0.93
- ICC_{BASDAI} = 0.94, 95% CI 0.92 to 0.96.

Schemmann and colleagues [33] found that the ICC coefficient of an app version of the iHOT-12 was 0.96, while the ICC coefficient of the paper version of this instrument was 0.92.

Concerning the administration of validated surveys to underage participants, Sun and colleagues [36, 37] found that the correlation coefficient of the FPS-R scores between their app and the paper version exceeded 0.60: Pearson's r > 0.90. Moreover, the mean differences were within +/- 0.24 (95% CI -1.57 to 1.97).

2.5.3.1.1.2. Time needed to complete surveys

Khraishi and colleagues [27, 28] found no difference between an app and paper in the mean time taken by participants to complete the HAQ: 95% CI -0.397 to 1.882, P = 0.193. However, Salaffi and colleagues [32] found a significant difference that favoured their app: MD = -2.80 minutes, 95% CI -3.19 to -2.41.

2.5.3.1.1.3. Acceptability to respondents

Concerning preference, Bush and colleagues [25] found that 73% of their participants indicated that they would prefer to use an iPhone if they were to complete the surveys again, and 76% of them said they would recommend the app-version of the survey (compared to 14% of participants who indicated no preference, and no one willing to recommend the paper version of the survey). Khraishi and colleagues [27, 28] found that 63% of their participants preferred the iPad® version of the survey; however, the authors did not report if the remaining 37% preferred the paper version of the survey or expressed no preference. Kim and colleagues [29], Newell and colleagues [31], and Salaffi and colleagues [32] observed that significantly more participants preferred the app version of the survey to its paper version: OR = 4.25, 95% CI 3.63 to 4.97; OR = 3.73, 95% CI 2.30 to 6.03; and OR = 543.32, 95% CI 30.79 to 9586.16. Furthermore, 58% of participants in Schemmann and colleagues [33] preferred the app version of the survey; however, the authors did not report the proportion of participants that preferred the paper version or that indicated no preference.

Regarding ease of use, Khraishi and colleagues [27, 28] found that 75% of participants rated an app as easier to use than paper. Newell and colleagues [31] observed no difference between delivery modes in participants' perceived difficulty in responding to a survey: $Mean_{App}$ = 1.31 (SD = 0.49), $Mean_{Paper}$ = 1.19 (0.58), t (100) = 1.75, P = 0.08, d = 0.18. Schemmann and colleagues [33] found that 42% of participants rated the app version of the survey as easier to use; however, the authors did not report if the remaining 58% of participants rated the paper version as easier to use, or indicated no difference. Using a five-point rating scale, Lamber and colleagues [30] observed no difference between delivery modes: $Mean_{Smartphone}$ = 4.55; $Mean_{Tablet}$ = 4.44; $Mean_{Paper}$ = 4.74; $Mean_{Laptop}$ = 4.60; Mann-Whitney tests, P > 0.05.

Kim and colleagues [29] observed that significantly more participants expressed their willingness to use the app version of the survey, compared to the paper version: OR = 2.56, 95% CI 2.20 to 2.97.

Concerning satisfaction, Lamber and colleagues [30] found no overall difference between satisfaction ratings on a five-point scale between delivery modes: $Mean_{Paper}$ = 4.42; $Mean_{Laptop}$

= 4.50; $Mean_{Smartphone}$ = 4.15; $Mean_{Tablet}$ = 4.44; Mann-Whitney test, P > 0.05. However, Newell and colleagues [31] observed a significant difference in the mean rating of a liking scale that favoured their app: $Mean_{App}$ = 4.03 (SD = 0.81); $Mean_{Paper}$ = 3.65 (SD = 0.86); t(97) = 3.66, P < 0.01, d = 0.45.

Regarding the effectiveness of the system informativeness, Lamber and colleagues found a potentially significant difference in mean scores that favoured the paper version of a survey (compared to its app version): $Mean_{Smartphone}$ = 4.25; $Mean_{Paper}$ = 4.66; Mann-Whitney tests, P = 0.05. Newell and colleagues found no difference in the mean scores of a clarity/confidence scale between an app and paper: $Mean_{App}$ = 4.43 (SD = 0.73); $Mean_{Paper}$ = 4.36 (SD = 0.68); t_S = 0.43, P_S = 0.67, d_S < 0.05.

Seventy-two percent of participants in Khraishi and colleagues [27, 28] reported feeling that the iPad® version of the survey took less time to complete than its paper version. In addition, 91% of participants in this study perceived the app version of the survey being more beneficial.

Newell and colleagues [31] observed no difference in the perceived ability to complete a survey between an app and paper: $Mean_{App}$ = 4.13 (SD = 0.85); $Mean_{Paper}$ = 4.13 (SD = 0.84); t (99) < 0.001, P > 0.99, d < 0.001.

Lastly, Sun and colleagues [36, 37] found that more participants (under the age of 18) preferred the app version of the FPS-R to the original instrument (P < 0.01).

2.5.3.1.2. App versus laptop computer

2.5.3.1.2.1. Data equivalence

Bush and colleagues [25] observed no difference between an app and a laptop computer in the mean scores of four symptom dimensions:

- MD_{PCL-C} = 2.90, 95% CI -2.87 to 8.67
- MD_{PHQ-9} = 0.10, 95% CI -1.99 to 2.19
- MD_{RSI-S} = 0.30, 95% CI -0.26 to 0.86
- $MD_{DRRI-US}$ = 0.80, 95% CI -5.32 to 6.92

Moreover, these authors found that the ICC coefficient for each symptom dimension exceeded 0.70:

- $ICC_{PCL\text{-}C}$ = 0.92, 95% CI 0.87 to 0.96
- $ICC_{PHQ\text{-}9}$ = 0.94, 95% CI 0.90 to 0.97
- $ICC_{RSI\text{-}S}$ = 0.87, 95% CI 0.80 to 0.94
- $ICC_{DRRI\text{-}US}$ = 0.93, 95% CI 0.90 to 0.97

2.5.3.1.2.2. Acceptability to respondents

Concerning preference, 13% of participants in Bush and colleagues [25] reported that they would use the laptop computer to complete the surveys again, while 73% reported they would use the smartphone. Similarly, 11% of participants indicated that they would recommend the laptop computer, compared to 76% of them who would recommend the smartphone.

Regarding ease of use, Lamber and colleagues [30] observed no difference between delivery modes on a five-point rating scale: $Mean_{Smartphone}$ = 4.55; $Mean_{Tablet}$ = 4.44; $Mean_{Paper}$ = 4.74; $Mean_{Laptop}$ = 4.60; Mann-Whitney tests, P > 0.05. Similarly, these authors did not find any difference between delivery modes on satisfaction ratings ($Mean_{Paper}$ = 4.42; $Mean_{Laptop}$ = 4.50; $Mean_{Smartphone}$ = 4.15; $Mean_{Tablet}$ = 4.44; Mann-Whitney tests, P > 0.05) or mean scores of system informativeness ($Mean_{Smartphone}$ = 4.25; $Mean_{Tablet}$ = 4.42; $Mean_{Laptop}$ = 4.46; Mann-Whitney tests, P > 0.05).

2.5.3.1.3. App versus tablet computer

2.5.3.1.3.1. Acceptability to respondents

Lamber and colleagues [30] observed no difference between delivery modes on a five-point rating scale assessing ease of use: $Mean_{Smartphone}$ = 4.55; $Mean_{Tablet}$ = 4.44; $Mean_{Paper}$ = 4.74; $Mean_{Laptop}$ = 4.60; Mann-Whitney tests, P > 0.05. Moreover, these authors found no difference between these delivery modes on a five-point satisfaction rating scale: $Mean_{Paper}$ = 4.42; $Mean_{Laptop}$ = 4.50; $Mean_{Smartphone}$ = 4.15; $Mean_{Tablet}$ = 4.44; Mann-Whitney tests, P > 0.05. Lastly, they observed no difference in the mean scores of system informativeness: $Mean_{Smartphone}$ = 4.25; $Mean_{Tablet}$ = 4.42; $Mean_{Laptop}$ = 4.46; Mann-Whitney tests, P > 0.05.

2.5.3.1.4. App versus PDA

2.5.3.1.4.1. Data equivalence

Brunger and colleagues [24] found that the correlation coefficients between an app and a PDA for each of five questions assessing appetite/satiety exceeded 0.60:

- How hungry do you feel? = 0.83, Standard Error of the Mean (SEM) = 0.04, range 0.27 to .94
- How full do you feel? = 0.76, SEM = 0.08, range -0.23 to 0.98
- How satiated are you? = 0.84, SEM = 0.05, range 0.17 to 0.98
- How strong is your desire to eat? = 0.85, SEM = 0.05, range 0.16 to 0.99
- How much could you eat? = 0.87, SEM = 0.04, range 0.73 to 0.99

2.5.3.1.5. App versus plastic ruler

2.5.3.1.5.1. Data equivalence

Sun and colleagues [36, 37] found that the correlation coefficients between CAS scores obtained through an app and those obtained using a plastic ruler exceeded 0.60 (Pearson's r > 0.90). However, the authors found that the mean app scores were higher than the plastic scores for all pairs of assessments: MD = 0.31, 95% CI -1.52 to 2.17, P < 0.03.

2.5.3.1.5.2. Acceptability to respondents

Sun and colleagues [36, 37] found that significantly more participants preferred the app version of the CAS to the original plastic instrument (P < 0.01).

2.5.3.2. Naturalistic settings

2.5.3.2.1. App versus paper

2.5.3.2.1.1. Data equivalence

Regarding non-validated surveys, Garcia-Palacios and colleagues [26] observed no difference between an app and paper in mean pain and fatigue scores:

- MD_{Pain} = 0.41, 95% CI -1.02 to 0.20
- $MD_{Fatigue}$ = 0.07, 95% CI -0.70 to 0.84

These authors also found that the correlation coefficients for each of these domains exceeded 0.60:

- $r_{Pain} = 0.79$, $P < 0.001$
- $r_{Fatigue} = 0.88$, $P < 0.001$

2.5.3.2.1.2. Data completeness

Garcia-Palacios and colleagues [26] found that an app resulted in significantly more complete records than paper: MD = 7.08, 95% CI 2.90 to 11.26. Similarly, these authors found that there were fewer incomplete records when the app was used: MD not estimable; t = 5.642, P < 0.01.

2.5.3.2.1.3. Adherence to survey administration protocols

Sigaud and colleagues [34] observed a difference in the rate of diary completion that favoured their app (P = 0.0398). Moreover, they found a difference in the adjusted mean of the intra-individual difference of completion rate that favoured the app (-19.5, 95% CI -38.10 to -0.90). Stomberg and colleagues [35] found that the mean number of entries on day one after surgery was lower in the app group than in the paper group (35 and 41 responses, respectively). On days two to four after surgery, the response rates were 100% in both groups. Lastly, on days five and six after surgery, the response rates in the app group were 69% on each day (participants completing the paper survey were followed up for four days only).

2.5.3.2.1.4. Acceptability to respondents

Regarding preference, Garcia-Palacios and colleagues [26] found no difference between an app and paper on a five-point scale: MD = -0.43, 95% CI -0.91 to 0.05. In addition, these authors observed a significant difference in the mean ratings given to an ease of use scale that favoured the app: MD= -0.62, 95% CI -0.91 to -0.33.

Sigaud and colleagues [34] found that 75% of their participants were willing to replace a paper diary with an app diary. They also observed that 79.2% of participants reported being satisfied with the app diary, compared to the paper diary. Garcia-Palacios and colleagues [26] observed

no difference between an app and paper in the system informativeness (defined here as perceived ease with which instructions could be followed): MD = 0.00, 95% CI -0.18 to 0.18.

Garcia-Palacios and colleagues [26] found a significant difference in the mean ratings of a five-point scale assessing the perceived time taken to answer the survey that favoured an app: MD= -0.32, 95% CI -0.62 to -0.02.

Lastly, Garcia-Palacios and colleagues [26] observed no difference between an app and paper in the mean ratings of a five-point scale assessing perceived usefulness: MD = -0.33, 95% CI -0.77 to 0.11.

2.5.3.2.2. App versus SMS

2.5.3.2.2.1. Data equivalence

Concerning the administration of validated surveys, Ainsworth and colleagues [23] observed no difference between an app and SMS in the mean scores of six symptom dimensions:

- $MD_{Hallucinations}$ = 0.20, 95% CI -0.82 to 1.22
- $MD_{Anxiety}$ = 0.70, 95% CI -0.28 to 1.68
- $MD_{Grandiosity}$ = 0.00, 95% CI -0.82 to 0.82
- $MD_{Delusions}$ = 0.10, 95% CI -0.58 to 0.78
- $MD_{Paranoia}$ = 0.30, 95% CI -0.69 to 1.29
- $MD_{Hopelessness}$ = 0.20, 95% CI -0.56 to 0.96

2.5.3.2.2.2. Data completeness

Ainsworth and colleagues [23] observed a significant difference in the mean number of daily entries between an app and SMS for days one, two, four, five and six of their sampling period:

- $MD_{Day\ 1}$ = 1.10, 95% CI 0.45 to 1.75
- $MD_{Day\ 2}$ = 1.40, 95% CI 0.84 to 1.96
- $MD_{Day\ 4}$ = 0.80, 95% CI 0.12 to 1.48
- $MD_{Day\ 5}$ = 1.00, 95% CI 0.24 to 1.76
- $MD_{Day\ 6}$ = 1.20, 95% CI 0.37 to 2.03

The difference in the mean number of entries on day three was not significant: MD = 0.40, 95% CI -0.37 to 1.17.

2.5.3.2.2.3. Time needed to complete surveys

Ainsworth and colleagues [23] found a significant difference in the mean time taken by participants to complete a survey that favoured their app: MD = -4.29 minutes, 95% CI -5.29 to -3.28.

2.5.3.2.2.4. Adherence to survey administration protocols

Ainsworth and colleagues [23] observed no difference between an app and SMS in the proportion of participants who completed at least one-third of all possible data collection points: OR = 1.84, 95% CI 0.39 to 8.77.

2.5.3.2.2.5. Acceptability to respondents

Ainsworth and colleagues [23] found that significantly more participants preferred an app to an SMS-based survey: OR = 14, 95% CI 3.19 to 61.36. In addition, these authors observed that significantly more participants found the app easier to use: OR = 12.14, 95% CI 3.03 to 48.67. Furthermore, Ainsworth and colleagues [23] observed no difference in the imagined length of time that participants would be willing to use either the app or SMS:

- $OR_{< 2 \text{ weeks}}$ = 0.64, 95% CI 0.10 to 4.20
- $OR_{2 \text{ to } 3 \text{ weeks}}$ = 1.00, 95% CI 0.32 to 3.15
- $OR_{3 \text{ to } 4 \text{ weeks}}$ = 0.17, 95% CI 0.02 to 1.54
- $OR_{4 \text{ to } 5 \text{ weeks}}$ = 3.29, 95% CI 0.32 to 34.08
- $OR_{5+ \text{ weeks}}$ = 1.90, 95% CI 0.52 to 6.97

These authors observed no difference in the mean overall scores of a quantitative feedback survey measuring the reactivity to the delivery mode and its successful integration into patients' daily routines: MD = -3.20, 95% CI -10.44 to 4.04.

2.6. Discussion

The aim of this chapter was to evaluate the impact that apps running on consumer mobile devices have on the quality of survey responses. Quality was assessed in relation to the survey error framework [4, 8]. Data equivalence, data accuracy and time needed to complete surveys were chosen as indicators of measurement error; data completeness, response rates, and adherence to survey administration protocols were chosen as indicators of representational errors. Acceptability to respondents was included as a systematic review outcome, as this is one of the critical success factors of any intervention.

One of the key sources of heterogeneity in this systematic review was the setting in which a study was conducted. The conditions afforded by different settings may determine which factors will be at operation during the survey completion process, or may cause the same factors to operate differently. Consequently, survey completion setting may influence the quality of survey responses.

Given the potential implications for practitioners, the findings of this systematic review were organised according to the main survey completion settings identified. In addition, findings were subsequently organised according to the type of comparison, as delivery mode effects tend to be mode-specific [1, 6, 9, 10, 12, 14].

Two main settings were identified: clinical and naturalistic. The former refers to locations where healthcare, social care, community care or research activities take place (e.g., hospital wards). In clinical settings, the connection between these activities and surveys is usually explicit, and patients can expect to complete one or more surveys. Moreover, practitioners are typically better able to standardise the conditions under which surveys are completed, and to implement quality assurance measures that require human involvement.

Naturalistic settings refer to the wide range of locations outside clinical or research environments such as home, work or outdoor spaces. In these settings, the conditions under which surveys are completed may vary both between patients (i.e., different patients completing the same survey but in different locations) and within patients (i.e., a patient completing multiple instances of the same survey in different locations). Moreover,

practitioners are not able to provide human input during survey completion; therefore, they might need to implement quality assurance measures that can successfully anticipate and mitigate any potential influence on survey responses. In addition, survey completion in naturalistic settings might involve the integration of a *medical instrument* into a patient's daily life, thus requiring the implementation of strategies for promoting effective engagement.

Outcome reporting varied between these settings. Data equivalence, time needed to complete surveys and acceptability to respondents were measured in both clinical and naturalistic settings. However, data completeness and adherence to survey administration protocols were reported by some studies conducted in naturalistic settings. This difference in outcome reporting highlights an important aspect of data collection: the need to select data quality dimensions and to implement data collection strategies that are relevant to the intended objective(s) and the characteristic of the setting in which surveys are completed. For example, studies conducted in naturalistic settings evaluated phenomena that required multiple data points over time (i.e., treatment adherence in patients with Haemophilia A, or the evaluation of post-surgical pain). In this context, it makes sense to implement longer and more frequent survey administration protocols, and assess patient adherence to these protocols. Similarly, data completeness becomes relevant, as practitioners might need a minimum number of data points to inform their decisions. These outcomes however, are not relevant in situations where patients are completing a survey once, just before seeing their clinician.

Concerning data equivalence, the evidence suggests that survey responses collected via smartphone or tablet apps are equivalent to responses collected via other delivery modes (i.e., paper, laptop computers, tablet computers pre-2007, PDA, and a plastic ruler). Regardless of setting, there were no differences between delivery modes in mean overall scores. Moreover, when measured, studies found that the ICC and other correlation coefficients exceeded 0.70 and 0.60, respectively. Although for one study, the wide confidence intervals for the mean difference in overall scores could indicate that the sample size was not large enough to derive a precise estimate of data equivalence between alternative delivery modes.

Moreover, the methodological characteristics of the included studies highlighted key issues around the choice of surveys. In both settings, surveys were administered in the context of an existing patient-clinician relationship. The purpose was to collect data that would inform clinical decisions. However, in clinical settings, only validated surveys were administered. Moreover, these surveys were administered only once. Studies in naturalistic settings administered predominantly non-validated surveys, following survey administration protocols of increased frequency and duration. The lack of validated surveys in naturalistic settings suggests that the conditions for survey completion and the survey administration protocols should resemble the conditions in which a survey was validated. Therefore, surveys that were validated for single administration should not be used for intensive data collection.

The findings concerning data equivalence also suggest that the adaptation of a survey for use with an app involves decisions about survey design. Studies using smartphones to administer surveys displayed one question per screen at any given time. This approach may be more appropriate for small screen devices. Moreover, findings suggest that the type of response scale might be an important factor when choosing a consumer mobile device. Regardless of data equivalence, Brunger and colleagues [24] observed that the length of a horizontal line influenced the sensitivity of participants' responses: ratings made on an iPad® showed subtle differences in perceived hunger, desire to eat and potential food intake than ratings made on a PDA.

Moreover, the findings suggest that the administration of surveys via an app also requires the implementation of data collection techniques that are compatible with survey administration protocols, and that can minimise the intrusion on a patient's daily routine. For example, studies conducted in naturalistic settings implemented diary and EMA techniques, which were more appropriate to the phenomena under evaluation. Moreover, in one of these studies the survey was divided into smaller question sets that were interspersed throughout the day (to avoid administering a lengthy survey in one session). In addition, considering typical usage patterns of consumer mobile devices might also help to reduce the likelihood of disrupting patients' routines.

Concerning the time needed to complete a survey, the findings are inconclusive. In clinical settings, apps were compared to paper with rheumatologic patients [27, 28, 32]. While one

study found that an app resulted in faster completion times, the other observed no difference between these delivery modes. This discrepancy could be the result of variation in app functionality. The functions implemented in one of these studies [32] (i.e., question branching, skip procedures, audio-recorded instructions with replay options) might have facilitated survey completion, and resulted in faster completion times. Alternatively, the discrepant findings could be the result of impaired usability due to poor app design strategies, particularly for patients with psoriatic and rheumatoid arthritis [27, 28]. For example, research has shown that the touch characteristics of patients with fine- and gross-motor impairments can be affected by interface design (e.g., button size) during digit entry tasks conducted on a touch screen [48]. Moreover, differences in the type and length of questions could also account for these findings. In a naturalistic setting, an app resulted in faster completion times than SMS [23]. These findings could also be explained in terms of app functionality or human-computer interaction factors. However, these possibilities were not explored by the authors.

Regarding acceptability to respondents, each study used their own bespoke definition. Defined as preference, this systematic review found that significantly more participants in clinical settings preferred an app to other delivery modes, whereas for participants in naturalistic settings the findings are inconclusive (although this could be the result of the types of comparisons made). Defined as ease of use, there were no clear differences between an app and other delivery modes when evaluated in clinical settings. In naturalistic settings however, significantly more participants found an app easier to use. In both settings, significantly more participants reported their willingness to use an app, compared to other delivery modes. Significantly more participants in naturalistic settings were satisfied with an app, while there was no clear difference between delivery modes in clinical settings. One study observed that the informativeness of a paper survey was superior to that of an app [30], whereas two other studies [26, 31] observed no difference in system informativeness between these delivery modes. The following definitions of acceptability to respondents were found to be relevant to longitudinal survey completion in naturalistic settings: maximum length of time that participants would be willing to use a delivery mode, and reactivity to a delivery mode and its successful integration into respondents' daily routines. Overall, these findings highlight the multi-faceted nature of acceptability, and suggest that it might be a

useful construct to identify usability issues that could influence the adoption of apps. More importantly, these findings highlight the need to examine if there is an association between facets of acceptability and personal characteristics within a target population (e.g., decreasing levels of perceived ease of use with age).

Data completeness and adherence to survey completion schedules were only assessed by some of the studies conducted in naturalistic settings. Two studies measuring data completeness found that an app resulted in significantly more complete records than paper [26], and in significantly more entries than with an SMS survey [23]. As with other systematic review outcomes, these findings could be the result of app functionality, human-computer interaction factors or additional interventions. For example, the implementation of alerts or reminders could have been reduced data incompleteness due to forgetfulness. These features are known to improve adherence to antiretroviral therapy in patients with human immunodeficiency virus (HIV) [49], or increase sunscreen use in at-risk groups [50]. However, these possibilities were not explored in these studies.

The findings concerning adherence to sampling protocols are inconclusive. Two studies [23, 35] observed no difference between an app, and SMS or paper on this outcome. However, another study [34] found that adherence to an app-based diary was better than for a paper diary. In addition to the factors mentioned previously (i.e., functionality, human computer interaction and additional interventions), the characteristics of the clinical population from which participants were recruited might also account for this discrepancy between studies. For example, the delivery mode might not affect respondents' adherence where there is an immediate, short-term goal (such as monitoring post-surgical pain). However, in long-term situations (e.g., self-management of a long-term condition such as diabetes) it might be important to consider strategies to promote patient engagement over time. Moreover, regardless of the delivery mode, regular self-monitoring might not be appropriate in all cases. For example, participants with schizophrenia in one of the included studies [23] reported mild negative reactivity to the continuous monitoring of their symptoms.

None of the included studies assessed data accuracy or response rates. The lack of studies reporting data accuracy might reflect the type of surveys administered for which there are no correct answers given the subjective nature of the attributes they assess. Moreover, response

rates are used to determine if respondents differ significantly from non-respondents. As such, it is a property of the sampling method and not of the survey. Apps were not used as a sampling method in any of the included studies. Instead, potential participants were approached and recruited directly as part of their routine care. Therefore, the potential role of apps for inviting and recruiting participants could not be assessed.

2.6.1. Limitations

As a delivery mode, apps can support the administration of both self-administered and interview-administered surveys. This chapter however, only evaluated the former. Interview-administered surveys were not considered here, as the presence of an interviewer is an independent influencing factor on the survey generation process [1, 4, 7, 15]. Moreover, this chapter considered situations where survey responses were used to inform clinician-led decisions. However, apps could support self-management or behaviour change activities where patients can use survey responses to inform their own care decisions. In addition, only structured responses collected through traditional responses scales were evaluated in this systematic review. Yet, apps and consumer mobile devices in general have enabled the collection of unstructured clinical data whose validity needs to be determined. Therefore, the role of apps in these situations must be explored separately.

Furthermore, this systematic review evaluated native apps running on smartphones or tablets computer that were released on or after 2007. Therefore, the following forms of technology were excluded: mobile self-administered surveys rendered on a web browser, feature phones with internet capabilities, alternative media for the delivery of questions and response options (e.g., SMS or pictures), manual or automated entry of data produced by medical devices with or without connectivity capabilities, wearable devices, and automated sensors. This decision was informed by the need to control for potential confounders such as differences in usability between platforms and devices, differences in the technical capabilities of devices, and issues of poor connectivity. Nonetheless, these technologies can play a key role in the collection of clinical data.

2.6.2. Conclusions

This chapter showed that, in the context of self-administered surveys, the choice of data quality domains must be relevant to a project's objective and data collection strategy (which includes survey completion settings and survey administration protocols). Concerning measurement error, the findings suggest that survey responses collected through apps are equivalent to those collected through other delivery modes. This appears to be the case in situations where the intended clinical application of a survey, its recommended frequency of administration, and the setting in which it was validated remain unchanged. There was no data on data accuracy, and the findings concerning the time needed to complete a survey were inconclusive. Concerning representational errors, there was no data on response rates; therefore, it was not possible to assess if apps can act as an effective sampling method. Moreover, although apps might improve data completeness, there was insufficient evidence to assess their impact on adherence to survey completion schedules. In addition, acceptability to respondents appears to be a useful construct to identify usability issues that might influence the successful adoption of apps.

This chapter also highlighted some of the factors that could result in delivery mode effects when using apps: survey design; app functionality; human-computer interaction factors; and patient characteristics, including their clinical condition(s). However, none of the included studies evaluated the potential influence of any of these factors. Such delivery mode effects are relevant because they could lead to diagnostic or therapeutic decisions that are not the result of an underlying clinical construct. Moreover, it has been suggested that mHealth has the potential to improve many aspects of healthcare delivery. However, the introduction of mobile technology into existing healthcare pathways is faced with implementation challenges that could prevent its successful adoption and potential benefits. Interestingly, none of the studies included in this systematic review assessed any of these challenges. Instead, these studies were conducted under highly-controlled conditions, which might not be representative of the clinical situations for which these apps were developed.

These findings informed the work that is presented in subsequent chapters, which was conducted in the context of antenatal care provided through the NHS. This context provided a relatively highly-structured patient-clinician relationship and a clinical purpose: the

identification of antenatal depression. Chapters 3 and 4 evaluated the feasibility of using a tablet computer app for depression screening in the waiting area of NHS antenatal clinics [54]. This corresponds to the clinical settings identified in this systematic review. For this study, survey layout was chosen as one components of survey design that could affect the responses given to two validated depression screening scales. Chapter 5 evaluated the feasibility of using a bespoke app running on patients' own smartphones for the longitudinal monitoring of mood and depression. This corresponds to the naturalistic settings identified in this systematic review. The main focus of this study was on the effect that a sampling protocol had on patient adherence. Both studies explored implementation challenges associated with the deployment of mobile technology in antenatal pathways.

References (Chapter 2)

1. Bowling, A. Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health (Oxford, England)*. 2005;27(3): 281-91. Available from: doi: 10.1093/pubmed/fdi031 [Accessed 28 of January 2017].
2. Boynton P, Greenhalgh T. Hands-on guide to questionnaire research: Selecting, designing, and developing your questionnaire. *BMJ*. 2004;328: 1312-1315. Available from: doi: https://doi.org/10.1136/bmj.328.7451.1312 [Accessed 28 January 2017].
3. Carter Y, Shaw S, Thomas C (eds.) *The Use and Design of Questionnaires*. Master Classes in Primary Care Research No 2. London, UK: The Royal College of General Practitioners; 2000.
4. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey Methodology*. 2nd Edition. New Jersey, US: Wiley; 2009.
5. Hosking JD, Newhouse MM, Bagniewska A, Hawkins BS. Data collection and transcription. *Controlled Clinical Trials*. 1995;16(2 Suppl): 66S-103S. Available from: doi: http://dx.doi.org/10.1016/0197-2456(94)00094-J [Accessed 28 January 2017].
6. Bowling A. *Research Methods in Health: Investigating Health and Health Services*. 3rd Edition. Maidenhead, Berkshire: McGraw Hill; 2009.
7. Streiner DL, Norman GR. *Health Measurement Scales: a practical guide to their development and use*. 4th Edition. Oxford, UK: Oxford University Press; 2008.
8. Lavrakas, PJ (ed.). *Encyclopedia of Survey Research Methods*. London, UK: SAGE Publications; 2008.
9. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. *Value IN Health*. 2008;11(2): 322-33. Available from: doi: 10.1111/j.1524-4733.2007.00231.x [Accessed 28 January 2017].
10. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, Lenderking WR, Cella D, Basch E, ISPOR ePRO Task Force. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. *Value in Health*. 2009;12(4): 419-29. Available from: doi: 10.1111/j.1524-4733.2008.00470.x [Accessed 28 January 2017].

11. Lampe AJ, Weiler JM. Data Capture From the Sponsors' and Investigators' Perspectives: Balancing Quality, Speed, and Cost. *Therapeutic Innovation & Regulatory Science*. 1998;32(4): 871-86. Available from: doi: https://doi.org/10.1177/009286159803200403 [Accessed 28 January 2017].
12. Lane SJ, Heddle NM, Arnold E, Walker I. A review of randomized controlled trials comparing the effectiveness of hand held computers with paper methods for data collection. *BMC Medical Informatics and Decision Making*. 2006;6:23.
13. Shih TH, Fan X. Comparing response rates in e-mail and paper surveys: a meta-analysis. *Educational Research Review*. 2009;4(1): 26-40. Available from: doi: 10.1016/j.edurev.2008.01.003 [Accessed 28 January 2017].
14. Fan W, Yan Z. Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*. 2010;26(2): 132-9. Available from: doi: 10.1016/j.chb.2009.10.015 [Accessed 28 January 2017].
15. Tourangeau R, Rips LJ, Rasinski K. *The Psychology of Survey Response*. New York, US: Cambridge University Press; 2000.
16. Wells T, Bailey JT, Link MW. Comparison of Smartphone and Online Computer Survey Administration. *Social Science Computer Review*. 2014;32(2): 238-255. Available from: doi: 10.1177/0894439313505829 [Accessed 28 January 2017].
17. Marcano Belisario JS, Jamsek J, Huckvale K, O'Donoghue J, Morrison C, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. Cochrane Database of Systematic Reviews. 2015;7: MR000042. Available from: doi: 10.1002/14651858.MR000042.pub2 [Accessed on 28 January 2017].
18. Manfreda KL, Bosnjak M, Berzelak J, Haas I, Vehovar V. Web surveys versus other survey modes - A meta-analysis comparing response rates. *International Journal of Market Research*. 2008;50(1): 79 - 104.
19. Gaggioli A, Pioggia G, Tartarisco G, Baldus G, Corda D, Cipresso P, Riva G. A mobile data collection platform for mental health research. *Personal and Ubiquitous Computing*. 2013;17(2): 241-251. Available from: doi: 10.1007/s00779-011-0465-2 [Accessed 28 January 2017]..
20. Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*. 2012;45(1): 184-98. Available from: doi: 10.1016/j.jbi.2011.08.017 [Accessed 28 January 2017].

21. Rodgers M, Sowden A, Petticrew M, Arai L, Roberts H. Testing Methodological Guidance on the Conduct of Narrative Synthesis in Systematic Reviews Effectiveness of Interventions to Promote Smoke Alarm Ownership and Function. *Evaluation*. 2009;15(1): 47-71.
22. Ainsworth J, Palmier-Claus JE, Machin M, Barrowclough C, Dunn G, Rogers A, Buchan I, Barkus E, Kapur S, Wykes T, Hopkins RS, Lewis S. A comparison of two delivery modalities of a mobile phone-based assessment for serious mental illness: native smartphone application vs text-messaging only implementations. *Journal of Medical Internet Research*. 2013;15(4): e60. Available from: doi: 10.2196/jmir.2328 [Accessed 29 January 2017].
23. Brunger L, Smith A, Re R, Wickham M, Philippides A, Watten P, Yeomans MR. Validation of an iPad visual analogue rating system for assessing appetite and satiety. *Appetite*. 2015;84: 259-63. Available from: doi: 10.1016/j.appet.2014.10.023 [Accessed 29 January 2017].
24. Bush NE, Skopp N, Smolenski D, Crumpton R, Fairall J. Behavioral screening measures delivered with a smartphone app: psychometric properties and user preference. *Journal of Nervous and Mental Disease*. 2013. 201(11): 991-5. Available from: doi: 10.1097/NMD.0000000000000039 [Accessed 29 January 2017].
25. Garcia-Palacios A, Herrero R, Belmonte MA, Castilla D, Guixeres J, Molinari G, Baños RM. Ecological momentary assessment for chronic pain in fibromyalgia using a smartphone: a randomized crossover study. *European Journal of Pain*. 2014;18(6): 862-72.
26. Khraishi M, Aslanov R. The use of an Apple iPad-based© health assessment questionnaire (HAQ) application in psoriatic and rheumatoid arthritis. *Dermatology and Therapy*. 2012;2(10): S58.
27. Khraishi M, Aslanov R, Ogunyemi B, Gu M, Lin L. Comparing touch screen Apple iPad-based© Health Assessment Questionnaire (HAQ) to the paper version in patients with inflammatory arthritis. *Journal of Rheumatology*. 2012;39(8): 1752.
28. Kim JH, Kwon SS, Shim SR, Sun HY, Ko YM, Chun DI, Yang WJ, Song YS. Validation and reliability of a smartphone application for the International Prostate Symptom Score questionnaire: a randomized repeated measures crossover study. Journal of Medical Internet Research. 2014;16(2): e38. Available from: doi: 10.2196/jmir.3042 [Accessed 29 January 2017].

29. Lamber P, Mitterer M, Napolitano L, Ricci F, Zini F. Surveying patients with smart devices. *Proceedings of the 25th International Symposium on Computer-Based Medical Systems*, 20-22 June 2012, Rome Italy. Available from: doi: 10.1109/CBMS.2012.6266349 [Accessed 29 January 2017].
30. Newell SM, Logan HL, Guo Y, Marks JG, Shepperd JA. Evaluating tablet computers as a survey tool in rural communities. *Journal of Rural Health*. 2015;31(1): 108-117. Available from: doi: 10.1111/jrh.12095 [Accessed 29 January 2017].
31. Salaffi F, Gasparini S, Ciapetti A, Gutierrez M, Grassi W. Usability of an innovative and interactive electronic system for collection of patient-reported data in axial spondyloarthritis: comparison with the traditional paper-administered format. *Rheumatology (Oxford)*. 2013;52(11): 2062-70. Available from: doi: 10.1093/rheumatology/ket276 [Accessed 29 January 2017].
32. Schemmann D., et al., Validation and patient acceptance of a touch tablet version of the iHOT-12 questionnaire. *Arthroscopy*. 2013;29(12 Suppl): e188.
33. Sigaud M, Horvais V, Chamouard V, Guillet B, Lambert T, Borel-Derlon A, Negrier C, Dolimier E, Masquelier AM. Comparison of paper diary and B-CoNect (telemetric smartphone application) at home treatment monitoring of severe hemophilia A patients. *Haemophilia*. 2014;20(Suppl. 3): 180.
34. Stomberg MW, Platon B, Widen A, Wallner I, Karlsson O. Health information: what can mobile phone assessments add? *Perspectives in Health Information Management*. 2012;9(Fall): 1d.
35. Sun T, West N, Ansermino M, von Baeyer C. PANDA: Evaluation of a smartphone-based perioperative pain assessment tool. Journal of Investigative Medicine. 2013;61(1): 115.
36. Sun, T., et al., PANDA: Evaluation of a smartphone-based peri-operative pain assessment tool. *Canadian Journal of Anesthesia*. 2013;60: S25.
37. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: dimensions and practical applications. *Health and Quality of Life Outcomes*. 2003;1: 20. Available from: doi: 10.1186/1477-7525-1-20 [Accessed 30 January 2017].
38. Barry MJ, Fowler FJ, O'Leary MP, Bruskewitz RC, Holtgrewe HL, Mebust WK, Cockett AT. The American Urological Association Symptom Index for Benign Prostatic Hyperplasia. The Journal of Urology. 2017;197(2 Suppl): S189-S197. Available from: doi: 10.1016/j.juro.2016.10.071 [Accessed 30 January 2017].

39. Barry MJ, Fowler FJ, O'Leary MP, Bruskewitz RC, Holtgrewe HL, Mebust WK, Cockett AT. The American Urological Association symptom index for benign prostatic hyperplasia. The Measurement Committee of the American Urological Association. Journal of Urology. 1992;148(5): 1549-57; discussion 1564.
40. EORTC Quality of Life. *EORTC QLQ - C30*. Available from: http://groups.eortc.be/qol/eortc-qlq-c30 [Accessed 01 February 2017].
41. Eaton WW, Muntaner C, Smith C, Tien A, Ybarra M. Center for epidemiologic studies depression scale: review and revision (CESD and CESD-R). In: Maruish ME (ed.) *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum; 2004.
42. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement*. 1977; 1(3): 383-401.
43. Higgins ET, Friedman RS, Harlow RE, Idson LC, Ayduk ON, Taylor A. Achievement orientations from subjective histories of success: promotion pride versus prevention pride. *European Journal of Social Psychology*. 2001. 31(1): 3-23. Available from: doi: 10.1002/ejsp.27 [Accessed 01 February 2017].
44. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *Journal of Rheumatology*. 1994;21(12): 2286-91.
45. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, Jenkinson T. A new approach to defining functional ability in ankylosing spondylitys: the development of the Bath Ankylosing Spondylitis Functional Index. *Journal of Rheumatology*. 1994;21(12): 2281-5.
46. Griffin DR, Parsons N, Mohtadi NG, Safran MR. A short version of the International Hip Outcome Tool (iHOT-12) for use in routine clinical practice. *Arthroscopy*. 2012;28(5): 611-6; quiz 616-8. Available from: doi: 10.1016/j.arthro.2012.02.027 [Accessed 01 February 2017].
47. Sesto ME, Irwin CB, Chen KB, Chourasia AO, Wiegmann DA. Effect of touch screen button size and spacing on touch characteristics of users with and without disabilities. *Human Factors*. 2012;54(3): 425-436.
48. Rodrigues R, Shet A, Antony J, Sidney K, Arumugam K, Krishnamurthy S, D'Souza G, DeCosta A. Supporting adherence to antiretroviral therapy with mobile phone reminders:

results from a cohort in South India. *PLoS One*. 2012;7(8): e40723. Available from: doi: 10.1371/journal.pone.0040723 [Accessed 01 February 2017].

49. Armstrong AW, Watson AJ, Makredes M, Frangos JE, Kimball AB, Kvedar JC. Text-message reminders to improve sunscreen use: a randomized, controlled trial using electronic monitoring. *Archives of Dermatology*. 2009;145(11): 1230-6. Available from: doi: 10.1001/archdermatol.2009.269 [Accessed 01 February 2017].
50. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Morrison C, Car J. Tablet computers for implementing NICE antenatal mental health guidelines: protocol of a feasibility study. *BMJ Open*. 2016;6(1): e009930. Available from: doi: 10.1136/bmjopen-2015-009930 [Accessed 01 February 2017].

3. Tablet computer apps for depression screening in clinical settings

3.1. Overview

Mobile devices may facilitate depression screening in the waiting area of antenatal clinics. However, this may present certain implementation challenges. This chapter describes a study that assessed the feasibility of using tablet computers to administer the Whooley questions and the EPDS to pregnant women attending NHS antenatal clinics in England. The primary outcomes were participants' responses to these depression screening scales, and whether they differed from expected prevalence rates in antenatal populations. Secondary outcomes included completion times, and requests for technical assistance. These outcomes were used as indicators of the potential impact that the deployment of mobile technology could have on the workload and schedules of antenatal clinics. Additional outcomes included smartphone and tablet computer ownership, breakoff rates, and the deployment challenges identified during the study. Moreover, controlled randomisation was used to compare two survey layouts in relation to the primary and secondary outcomes.

The proportion of mothers-to-be screening positive for depression in this study was in agreement with expected prevalence rates for this population. This finding suggests that antenatal depression screening using tablet computers might not result in a disproportionate number of patients screening positive, which would add to the existing workload of antenatal services. In addition, the results showed that survey layout did not affect the responses given to the Whooley questions, the median EPDS scores, the number of participants at increased risk of self-harm, and the number of participants asking for technical assistance. However, there was a significant association between survey layout and the proportion of participants at each scoring interval of the EPDS. This association must be explored further to understand its full implications for the provision of antenatal mental health care. Furthermore, a scrolling layout resulted in faster completion times than a paging layout. However, the significance of this difference is yet to be determined.

Tablet computers can be used for depression screening in the waiting area of antenatal clinics, as long as this practice is sensitive to clinical workflows, and technology-related issues such as connectivity and security.

3.2. Rationale

The aim of this chapter is to apply the findings from the systematic review outlined in Chapter 2 to the assessment of depression during pregnancy in the context on antenatal care provided through the NHS. Since chapter 2 highlighted that survey responses collected via a mobile application are equivalent to those collected through traditional delivery modes, a decision was made not to compare an app to paper. Instead, the focus is on two of the other main lessons from Chapter 2, which have not been explored in the healthcare literature: (i) the importance of setting as a determinant of the factors that influence survey completion and survey responses; and (ii) the impact of survey design choices on survey responses.

Therefore, this chapter presents the results of a study of the feasibility of using a tablet computer app in the waiting area of NHS antenatal clinics to implement NICE recommendations for the recognition of depression during pregnancy. This corresponded to the clinical settings discussed in Chapter 2, and provided an opportunity to explore the implementation challenges associated with the deployment of mobile technology in clinical practice. In addition, this study explored the impact that survey layout (as a component of survey design) had on the responses given to two depression screening scales.

3.3. Background

The role of mobile technology for the assessment and screening of mental health disorders has received much attention in recent years [1-5]. The research outlined in Chapter 2 has shown that the administration of surveys via an app does not affect the overall equivalence of the responses collected, compared to other delivery modes [6]. Since case-finding in non-mental health settings (e.g., primary care or antenatal care) relies on the use of validated surveys, smartphones and tablet computers may facilitate the practical implementation of mental health screening in these settings. For patients, these devices could make the completion of screening scales more convenient, while increasing their perceived sense of confidentiality. For clinicians, these devices could enable rapid and convenient access to patients' risk of a given mental health condition.

Mental health screening through sconsumer mobile devices could be particularly beneficial for maternal mental health. The term perinatal is used to cover the time between pregnancy and the first year after birth. Mental health during this period encompasses several conditions, the most common of which are depression and anxiety. However, this project focused only on depression that occurs during pregnancy (or antenatal depression). This decision was informed by practical reasons (i.e., project timelines and resources), as well as by public health reasons.

From a public health perspective, antenatal depression has been estimated to affect between 7% and 12% of pregnant women [7, 8]. Period prevalence estimates suggest that as many as 12.7% of women could experience an episode of major depression at any point during their pregnancy [7, 8]. If left untreated, this condition can have severe and long-lasting consequences for mothers, their children, and their families [9-15]. Pregnant women struggling with depression are more likely to engage in unhealthy behaviours such as poor diet, substance abuse, and failure to attend antenatal care appointments [9, 12]. In addition, these women are at increased risk of self-harm, and of postnatal depression (including more severe forms of this condition) [9, 12]. Furthermore, untreated antenatal depression can affect foetal development, and has been identified as an independent risk factor for a child's behavioural, cognitive, and emotional development [10-13]. Although there is no difference in depression incidence or prevalence between pregnant and non-pregnant women, research indicates that diagnosis and treatment rates are lower during pregnancy [16]. It has been estimated that up to three-quarters of pregnant women meeting diagnostic criteria for depression are not identified, and that only one in ten access appropriate treatment [17]. Furthermore, a recent report has identified important gaps in mental health care provision during pregnancy in England [14].

Some of the barriers to the recognition of antenatal depression include difficulties in differentiating depressive symptoms from expected mood and somatic changes during pregnancy; stigma; lack of reassurance that mental health care is a normal part of antenatal care; characteristics of healthcare providers; configuration of health services; insufficient consultation time; and the cost-effectiveness of screening practices [16-18]. As with other

mental health conditions, mobile technology could help address some of the practical barriers. Particularly, given the configuration of antenatal care in England.

NICE recommends that women with uncomplicated pregnancies should attend between 7 and 10 antenatal appointments, depending on their previous obstetric history [19]. These repeated interactions between a pregnant woman and her care provider offer an opportunity for the timely administration of depression screening scales through mobile devices. Moreover, these scales could be implemented in the waiting area of antenatal clinics. Research has shown that patients who complete surveys whilst waiting for their appointment tend to report reduced frustration levels [20]. Gathering data beforehand could also trigger meaningful discussions during the consultation that patients may have trouble raising directly [20]. Lastly, this could facilitate the monitoring of disease progression and treatment response, and help to ensure that important issues are not overlooked by the clinical team [20]. However, the use of mobile technology for screening in clinical settings can present certain implementation challenges, of which this project focuses on survey design and deployment.

Findings from the survey methodology literature have shown that survey layout can influence respondents' behaviour, a phenomenon known as survey layout effect [21, 22]. Survey layout refers to the number of questions that are presented on a single screen at any given time, and includes scrolling layout and paging layout. The former refers to the presentation of all questions and response options on a single screen. Therefore, respondents need to scroll vertically to answer all questions. A paging layout on the other hand, refers to the presentation of one question per screen at any given time. In this case, respondents must navigate back and forth between multiple screens to complete a survey. Research has found that scrolling layouts result in lower breakoff and item non-response rates, fewer technical problems, faster completion times, and higher subjective ratings than paging layouts [21]. Although these findings concern volunteer activities surveys, they could have clinical implications.

A survey layout effect on clinical scales has the potential to introduce artefacts into the clinical decision-making process. Depending on its size and direction, a survey layout effect on the subjective ratings given to a depression screening scale could influence a patient's result. This

could lead to patients undergoing unnecessary diagnostic pathways due to false positive results, or in diagnostic delays due to false negative results. In addition, there could be implications for clinicians in terms of increased workload due to false positive results, or liability due to false negatives.

The findings from the survey methodology literature also raise practical implications for the deployment of mobile technology in clinical settings. The potential for increased workload of clinical staff, training and capacity needs, location and privacy of screening, network issues, and responsibility for technology are some of the perceived barriers to the success of mobile-based antenatal depression screening [5, 23]. Indicators from survey methodology studies (e.g., completion times, breakoff rates, technical problems) may allow practitioners to capture useful information about the real impact of these perceived barriers, giving them the opportunity to address them.

Therefore, assessing the feasibility of mobile technology for antenatal depression screening is warranted. In doing so, it is important to consider potential survey layout effects that could influence the screening process, as well as the challenges that practitioners might face when deploying mobile technology in their clinics. This chapter aims to provide such assessment.

3.4. Methods

This study received favourable opinion from the South East Coast – Surrey National Research Ethics Committee on 09 July 2015, under the reference number 15/LO/0977. The methods described in this chapter were published as a study protocol and as a research paper in peer reviewed journals [24, 25].

3.4.1. Aim

This main aim of this study was to assess the feasibility of using tablet computers (Apple® iPads®) in the waiting area of antenatal clinics for implementing NICE guideline recommendations for recognising antenatal depression [26]. The primary outcomes were participants' responses to the Whooley questions and the EPDS, and whether these differed from prevalence depression rates reported in the literature.

Secondary outcomes included completion time and the number of participants requesting technical assistance. These outcomes were chosen as indicators of the impact that the deployment of mobile technology could have on the workload and schedules of clinics.

This study explored differences between two survey layouts (i.e., scrolling and paging) across all primary and secondary outcomes.

Additional outcomes included smartphone and tablet computer ownership and breakoff rates. The former provided a useful indication of participants' familiarity with mobile technology, and of the potential reach of future mobile health interventions. Breakoff rates showed the incompatibility between research procedures and clinic timings. Lastly, the deployment challenges identified during this study were recorded and reported here.

Item non-response rates were not measured, as validation procedures did not allow participants to leave questions unanswered.

3.4.2. Study design

This study assessed the feasibility of using tablet computers to administer a socio-demographic survey, the Whooley questions and the EPDS to 530 pregnant women attending antenatal clinics in 14 NHS facilities across England. In addition, participants were randomly allocated to one of two layout versions of these surveys:

- Paging layout: in which one question per screen was displayed on the iPad® screen at any given time
- Scrolling layout: in which one survey per screen (regardless of the number of questions) was displayed on the iPad® screen at any given time

3.4.3. Sample size

At least 30 events (defined as pregnant women scoring 10 points or more on the EPDS) were needed to calculate differences in the outcomes of interest. This required a sample of 250 participants, assuming a prevalence of antenatal depression of 12%. Therefore, the aim was

to recruit at least 300 participants to account for eventualities such as dropouts or a higher prevalence rate in the target population.

3.4.4. Participants and participant recruitment

Adult (18 years or older) pregnant women of any gestational age, pregnancy history and parity were recruited from participating NHS antenatal clinics (in general practices, community services or secondary care centres).

The main focus of this feasibility study was on the role of mobile technology for identifying pregnant women who were not known to mental health services and could benefit from additional support. For this reason, women who had been diagnosed with depression or generalised anxiety disorder in the past 12 months (from the moment they were approached by the research team), or who were currently receiving treatment for any of these disorders were excluded.

In addition, women who did not feel comfortable reading and writing in English were also excluded. Although one of the perceived advantages of mobile technology is that it can facilitate the administration of surveys in multiple languages, it would still have been necessary to identify multiple validated translations of the Whooley questions and the EPDS. This would have been impractical from a resourcing perspective given the regional variation in the ethnic composition of antenatal populations. In certain regions, the exclusion of certain languages would have been inevitable due to the lack of a validated translation of the Whooley question and of the EPDS.

See Table 2 for participant inclusion and exclusion criteria. Unfortunately, the clinics from which participants were recruited did not keep a record of the number of women that were excluded on the basis of these criteria.

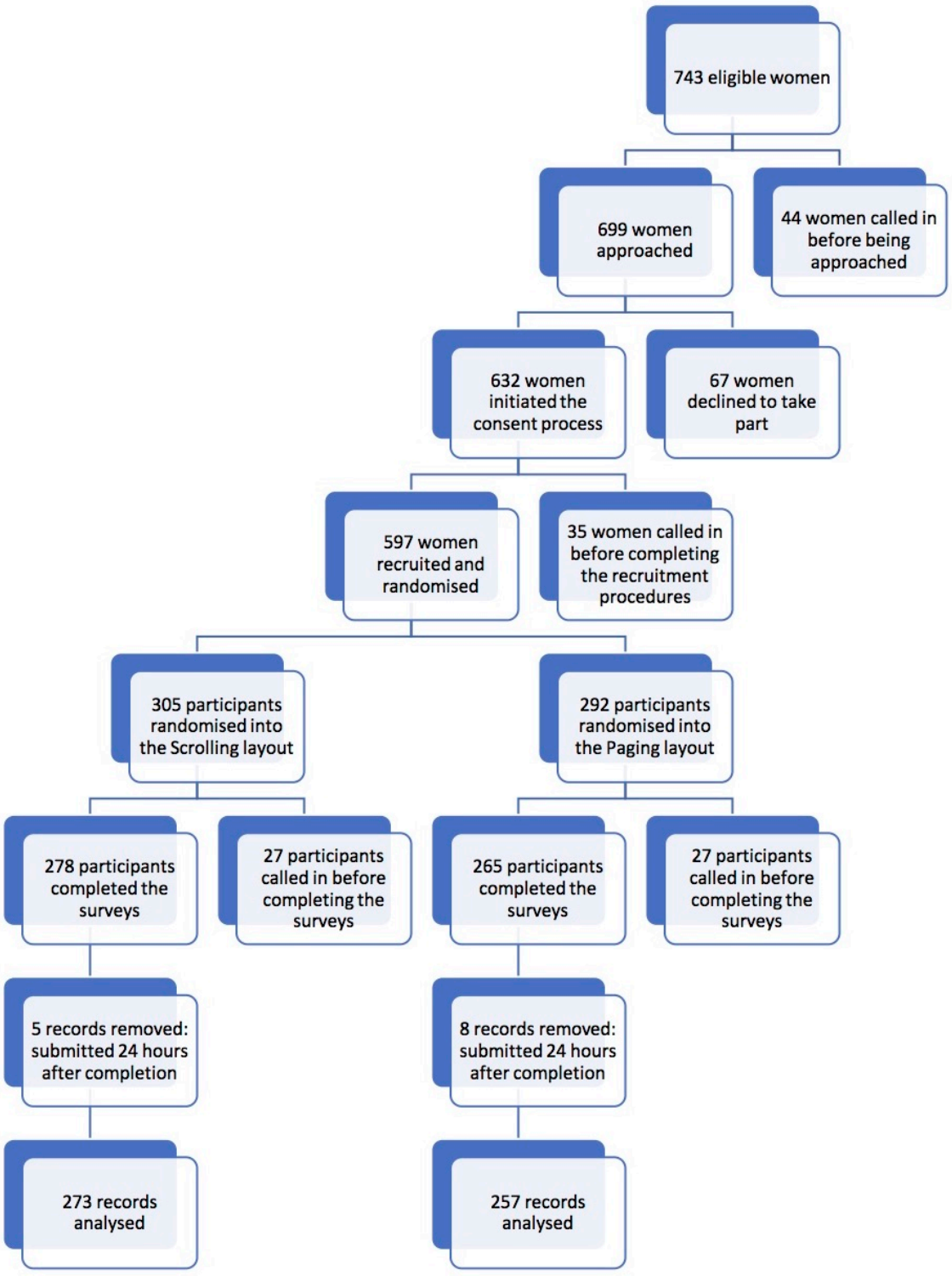
Table 2. Participant inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
18 years or older	Self-disclosed mental health problem at the time of being approached about participation, or in the previous 12 months
Pregnancy (of any gestational age and parity)	Currently receiving treatment for any affective disorder
Attending a participating NHS antenatal clinic	Not comfortable reading and writing in English

Recruitment took place between October 2015 and May 2016 using an opportunistic approach. Pregnant women attending participating antenatal clinics were approached in the waiting area by a research midwife or a clinical studies officer (CSO) with information about the study and a copy of the participant information leaflet (PIL). Those consenting to take part in the study were then asked to complete an informed consent form and, subsequently, to complete the study procedures in the clinic's waiting area before their appointment. Participants were made aware that they could ask for help at any point while completing the surveys, and that all questions should be directed to a member of the research team (not to the clinical care team). Only technical assistance and clarification of questions were offered. Research midwives and CSOs did not read the questions out loud to the participants, or answered questions on their behalf.

As mentioned previously, the number of women who were not eligible to take part in this study was not recorded. Therefore, only the study flow information for eligible women is available. Overall, 743 eligible pregnant women were identified. Of these, 699 were approached with information about the study (as 44 were called in before they could be approached). Of these, 632 agreed to take part in the study and initiated the consent process. Therefore, the recruitment rate in relation to the total number of eligible participants (N = 743) was 85.10%. In relation to the total number of eligible women who were approached (N = 699), the recruitment rate was 90.41%. See Figure 6 for the study flow.

Figure 6. Study flow



Of the 632 women who initiated the consent process, 35 were called in before they could complete the informed consent form. Therefore, 597 participants were recruited into the study. Of these, 54 were called in (27 participants from the Scrolling group and 27 participants from the Paging group) before completing the surveys. Therefore, the completion rate was 90.95% (543 participants out of the 597 who were initially randomised).

3.4.5. Technology and materials

The surveys were implemented using Snap® survey software [27] and administered through Snap® Mobile app (versions 4.0.30 to 4.0.32) for Apple® iPhone Operating System (iOS™) [28] running on Apple® iPad Air® tablet computers. Responses were stored in Snap® WebHost [29]. These solutions supported the required survey layouts and offline data collection. The latter was an important requirement due to the variability in internet access across NHS facilities.

3.4.6. Surveys

3.4.6.1. Socio-demographic survey

A bespoke, non-validated socio-demographic survey was used to assess participants' age group, ethnicity, relationship status, employment status, level of education, smartphone and tablet ownership, pregnancy history and parity, pregnancy trimester, and previous history of depression (Appendix 32).

3.4.6.2. Validated, depression screening scales

Following the recommendations of the Antenatal and Postnatal Mental Health NICE guideline [26], the Whooley questions and the EPDS were administered to study participants.

3.4.6.2.1. Whooley questions

The Whooley questions consist of a 2-item survey evaluating the presence of depressed mood and anhedonia over the past month [30] (Appendix 33). Patients are required to answer *Yes* or *No* to each question. According to NICE recommendations, an affirmative answer to either question should be followed up by further assessment using a validated screening tool [26].

3.4.6.2.2. Edinburgh Postnatal Depression Scale

The EPDS is a 10-item self-administered survey that was originally developed to screen for postnatal depression [31]. Since then, this instrument has been validated for use in the perinatal period and for use in community and clinical settings. The EPDS assesses feelings of guilt, sleep disturbance, anhedonia and suicidal ideation that have been present for the past 7 days (Appendix 34). Each question is scored on a 4-point scale ranging from 0 to 3 points. An overall score is generated by the sum of these responses.

Variability in the diagnostic accuracy of different EPDS scores has been reported in the literature. Typically, the following thresholds are used:

- Ten points: possible depression
- Thirteen points: probable depression

A threshold of 15 points on the EPDS has also been proposed as suggestive of antenatal depression. However, this study implemented a threshold of 13 points or more as this is compatible with the care pathways of the antenatal clinics that participated in this study. Moreover, special attention was given to question 10, as it deals with thoughts of self-harm.

The EPDS represents a valid and reliable tool for identifying women who are at risk of depression, both during pregnancy and after childbirth. This instrument is also sensitive to changes in the severity of depression over time [31].

Although NICE guidelines recommend using a two-staged approach where the EPDS is administered only if there is an affirmative answer to at least one Whooley question, all the participants in this study completed the EPDS regardless of their answers to the Whooley questions.

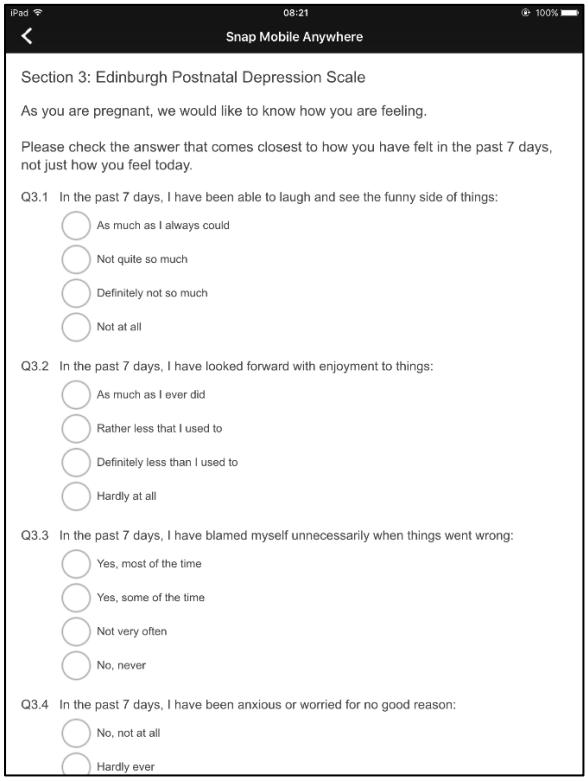
3.4.7. Survey layout

Two survey layouts of the socio-demographic survey, the Whooley questions and the EPDS were implemented: scrolling and paging. They differed in the number of questions that were presented on a single screen at any given time.

3.4.7.1. Scrolling layout

In this survey layout, each survey was presented on a single screen (see Figure 7). Therefore, participants had to scroll vertically to answer all the questions. Participants could navigate vertically and between surveys, and to modify their answers to any questions before submitting their final answers. After pressing the submit button, participants were asked to return the tablet computer to the research midwife or CSO. The latter was then presented with a summary of the participant's results. This information was used to complete a summary letter, which was given to a midwife or consultant in time for the consultation. Validation rules were applied with the app, which prevented participants from leaving any questions unanswered.

Figure 7. Scrolling layout through SNAP® surveys (screenshot)

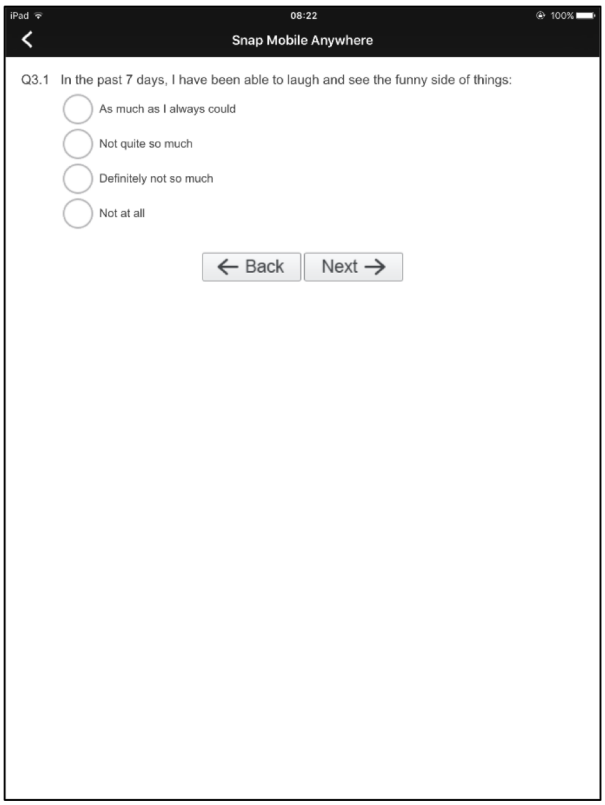


3.4.7.2. Paging layout

In this survey layout, only one question was displayed on the screen at any given time (see Figure 8). Therefore, participants were required to navigate back and forth between questions to complete all the surveys. Moreover, they could modify their answers before

pressing the submit button. The remaining study procedures were identical as those in the scrolling layout.

Figure 8. Paging layout through SNAP® surveys (screenshot)



3.4.8. Study procedures

3.4.8.1. Randomisation

Block randomisation (with blocks of 4) was used to allocate participants to one of two types of survey layout: paging or scrolling. For each participating site, an independent allocation sequence was generated using Stata V.13.0 [32]. These sequences were printed and placed inside opaque, numbered envelopes that were then sealed and distributed to each site. Research midwives and CSOs in charge of participant recruitment and administering the study procedures were not involved in allocation sequence generation. Envelopes were used sequentially and opened once a participant had completed the informed consent form. The contents of the envelop informed the research midwife or CSO about the version of the surveys they should activate; at this point, they handed over the tablet computer to the participant. If a participant was called in before completing the surveys, her record on the tablet was cancelled (and deleted) and the content of that randomisation envelope discarded.

3.4.8.2. Duty of care

The study procedures were completed in the waiting area, before a pregnant woman went into her appointment. Her results on the Whooley questions and the EPDS were summarised in a letter and communicated to the relevant midwife or consultant in time for the appointment. Those women with overall EPDS scores above 10 or 13 points, or scoring one point or more on question 10 of the EPDS (regardless of their overall score), received further assessment in accordance with existing local care pathways. However, resourcing issues prevented the research team from examining how the midwives or consultants used these letter summary letters in the consultation.

3.4.9. Statistical analyses

Descriptive statistics were presented for all primary and secondary outcomes.

The survey layouts were compared in relation to participants' responses to the Whooley questions and the EPDS:

- Proportion of participants answering Yes to one or more Whooley question(s)
- Median EPDS scores
- Proportion of participants scoring at each interval of the EPDS (i.e., 0 – 9 points, 10 – 12 points, and 13 points or more)
- Proportion of participants scoring 1 point or more on question 10 of the EPDS.

The survey layouts were also compared in relation to the following outcomes:

- Overall median time (in seconds) taken by participants to complete all the surveys
- Proportion of participants requesting technical assistance.

In addition, smartphone and tablet ownership across all participants were reported.

A Chi-squared test was used to compare proportions of participants between the scrolling and paging layouts. A Wilcoxon rank-sum test was used to compare EPDS scores and completion times, as these outcomes were not normally distributed.

Statistical analyses were conducted in R version 3.3.1 [33].

3.5. Results

The results presented in this section were published as a research paper in a peer-reviewed journal [25].

Participants were recruited from 14 localities across England (Blackburn, Bolton, Burton, Bury, East Midlands, London, Huntingdon, North East, Oldham, Rochdale, Shrewsbury, Stoke-on-Trent, Telford, and Wigan). Figure 6 shows the number of participants at each stage of the study. Of 597 pregnant women who consented to participate, 543 completed the study procedures. However, 13 records were removed as they were submitted more than 24 hours after completion (which affected the completion time registered by Snap® Mobile). Therefore, 530 records were included in the final statistical analysis: 273 from the scrolling group and 257 from the paging group.

Table 3 provides a summary of participants' socio-demographic characteristics, and shows that the two experimental groups were well-balanced in terms of the socio-demographic characteristics of the participants.

Table 3. Participants' socio-demographic characteristics

		Scrolling n (% out of 273)	Paging n (% out of 257)
Recruitment Setting	***General Practice***	59 (21.61%)	56 (21.79%)
	Community	14 (5.13%)	12 (4.67%)
	Hospital	200 (73.26%)	189 (73.54%)
Age group	***18 – 22 years***	15 (5.49%)	11 (4.28%)
	23 – 27 years	51 (18.68%)	34 (13.23%)
	28 – 32 years	90 (32.97%)	108 (42.02%)
	33 – 37 years	83 (30.40%)	81 (31.52%)
	38+ years	34 (12.45%)	23 (8.95%)
Ethnic Background	***White***	173 (63.37%)	182 (70.82%)
	Mixed	9 (3.30%)	7 (2.72%)
	Asian	64 (23.44%)	45 (17.51%)
	Black	14 (5.13%)	14 (5.45%)
	Other ethnic group	12 (4.40%)	9 (3.50%)
	Not disclosed	1 (0.37%)	0 (0.00%)
Relationship status	***Single***	17 (6.23%)	13 (5.06%)
	Married/Civil Partnership	180 (65.93%)	168 (65.37%)
	Cohabiting	71 (26.01%)	67 (26.07%)
	Divorced/Widowed/Separated	2 (0.73%)	3 (1.17%)
	Not disclosed	3 (1.10%)	6 (2.33%)
Employment status	***Full-time employment***	135 (49.45%)	141 (54.86%)
	Part-time employment	52 (19.05%)	49 (19.07%)
	Self-employed	25 (9.16%)	19 (7.39%)
	Seeking employment	6 (2.20%)	13 (5.06%)
	Unemployed	49 (17.95%)	34 (13.23)
	Disability	6 (2.20%)	1 (0.39)
Highest educational achievement	***Postgraduate degree***	56 (20.51%)	52 (20.23%)
	University degree	107 (39.19%)	94 (36.58%)
	College degree or below	95 (34.80%)	97 (37.74%)
	None	8 (2.93%)	5 (1.95%)
	Other qualification	7 (2.56%)	9 (3.50%)

3.5.1. Participants' responses to the Whooley questions and the EPDS

3.5.1.1. Whooley questions

Overall, 122 (23.02%) participants answered Yes to at least one Whooley question and 408 (76.98%) answered No to both questions (Table 4). A chi-squared test of independence was used to assess the relationship between participants' answers to the Whooley questions (i.e., the number of participants answering Yes to at least one questions and those answering No to both questions) and survey layout (i.e., scrolling and paging). The relationship between these two variables was non-significant, χ^2 (1, N=530) = 0.0049718, p = 0.944.

Table 4. Number of participants according to their answers to the Whooley questions

	***Yes* to at least one question** **n (row %)**	***No* to both questions** **n (row %)**
Scrolling (N = 273)	62 (22.71%)	211 (77.29%)
Paging (N = 257)	60 (23.35%)	197 (76.65%)
Total (N = 530)	**122 (23.02%)**	**408 (76.98%)**

3.5.1.2. Edinburgh Postnatal Depression Scale

Figure 9 shows the distribution of EPDS scores in this sample of participants. The distribution is right-skewed, with a skewness coefficient of 1.240029 that was shown to be significantly different from 0 ($p < 0.001$), thus suggesting non-normality. Moreover, a Shapiro-Wilk normality test also suggests that the EPDS scores are not normally distributed: W = 0.87806, $p < 0.001$.

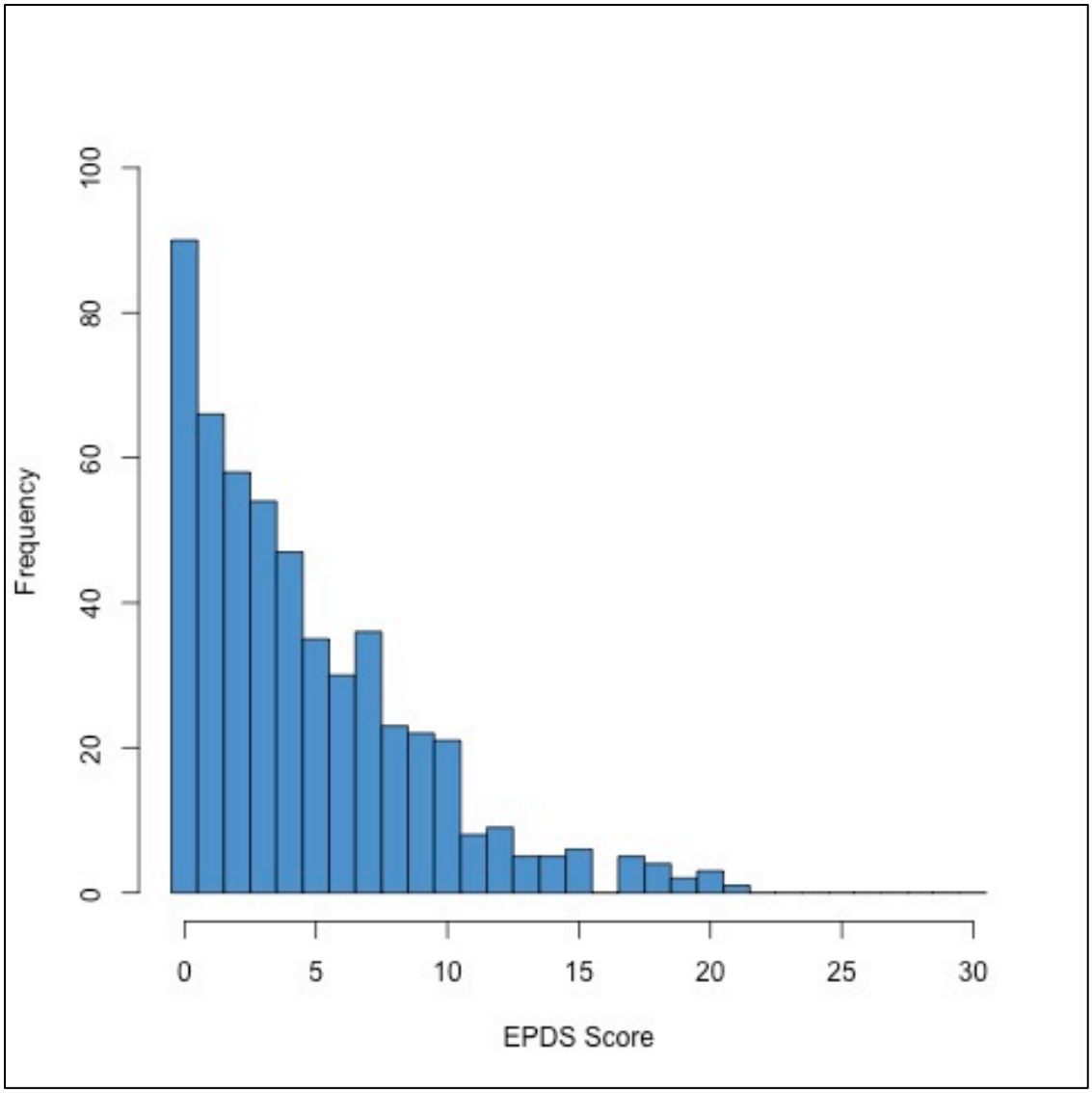


Figure 9. Distribution of EPDS scores in the sample of participants

Table 5 shows a statistical summary of the EPDS scores in the overall sample of participants, as well as by survey layout allocation. The maximum EPDS score in the *Scrolling* group was 21 points, whereas in the *Paging* group it was 17 points. A Wilcoxon rank-sum test was used to compare the median EPDS scores between both survey layouts, and no difference was found between them: W = 35755, p = 0.700.

Table 5. Participants' overall mean and median EPDS scores, and by survey layout

	Mean EPDS score (SD)	Median EPDS score	Minimum EPDS score	Maximum EPDS score	First quartile	Third quartile	IQR
Scrolling (N = 273)	4.72 (4.82)	3	0	21	1	7	6
Paging (N = 257)	4.43 (3.80)	3	0	17	1	7	6
Total (N = 530)	**4.58 (4.35)**	**3**	**0**	**21**	**1**	**7**	**6**

Table 6 shows the number of participants at each EPDS scoring interval, as well as by survey layout allocation. A chi-squared test of independence was used to examine the relation between EPDS scoring intervals and survey layout. The relation between these variables was significant, χ^2 (2, N = 530) = 9.6779, p = 0.008. This prompted the inclusion of the standardised residuals in Table 6. A visual inspection of these residuals revealed that the observed frequency of participants scoring 13 points or more on the EPDS in the *Scrolling* group exceeded the expected frequency by more than the critical value of 1.96 for a significance level of 0.05.

Table 6. Overall number of participants at each EPDS scoring interval, and by survey layout

		0-9 points n (row %)[1]	10-12 points n (row %)[1]	13+ points n (row %)[1]
Scrolling (N = 273)	Observed frequency	236 (86.45%)	14 (5.13%)	23 (8.42%)
	Expected frequency	237.46	19.57	15.97
	Standardised residuals	-0.38	-1.88	2.60
Paging (N = 257)	Observed frequency	225 (87.55%)	24 (9.34%)	8 (3.11%)
	Expected frequency	223.54	18.43	15.03
	Standardised residuals	0.38	1.88	-2.60
Total (N = 530)		**461 (86.98%)**	**38 (7.17%)**	**31 (5.85%)**

[1] Row % for observed frequency only and *Total* row

Table 7 shows the overall number of participants according to their score on question 10 of the EPDS, as well as by survey layout allocation. A chi-squared test of independence was used to examine the relation between participants' scores on question 10 of the EPDS and survey layout (i.e., Scrolling and Paging). The relation between these variables was non-significant χ^2 (1, N = 530) = 0.010247, p = 0.919.

Table 7. Number of participants according to their answers to question 10 of the EPDS, and by survey layout

	0 points **n (row %)**	**1+ points** **n (row %)**
Scrolling (N = 273)	268 (98.17%)	5 (1.83%)
Paging (N = 257)	251 (97.67%)	6 (2.33%)
Total (N = 530)	**519 (97.92%)**	**11 (2.08%)**

3.5.2. Completion time and requests for technical assistance

3.5.2.1. Completion time

Participants who requested technical help from the local research team during survey completion were excluded from the calculations concerning completion time. Therefore, data from 495 participants were included in this analysis: 254 from the scrolling group and 241 participants from the paging group.

Overall, 75% of participants completed the study procedures in 7 minutes 16 seconds or less, and half of them in 5 minutes 9 seconds or less. These completion times included the baseline demographic data survey, and are thus an overestimation of the real duration of the screening process (i.e., Whooley questions and EPDS). To give a sense of the overestimation, the Whooley questions and the EPDS were trialled with colleagues at Imperial College London, which yielded estimates of 2 minutes 28 seconds. A Wilcoxon rank-sum test revealed that participants in the *Scrolling* group were faster at completing the surveys (Median = 285.5 seconds) than participants in the *Paging* group (Median = 333 seconds): W = 27014, p = 0.024.

3.5.2.2. Requests for technical assistance

Overall, 35 participants made requests for technical assistance. A chi-squared test of independence was used to explore the relation between the number of participants asking for technical assistance and survey layout (i.e., Scrolling and Paging) (Table 8). The relation between these variables was non-significant χ^2 (1, N = 530) = 0.027251, p = 0.869.

Table 8. Overall number of participants requesting technical assistance, and by survey layout

	No requests made n (row %)	**Requests made n (row %)**
Scrolling (N = 273)	254 (93.04%)	19 (6.96%)
Paging (N = 257)	241 (93.77%)	16 (6.23%)
Total (N = 530)	**495 (93.40%)**	**35 (6.60%)**

No participant made more than one request for technical assistance. Table 9 provides a summary of the types of requests made.

Table 9. Number and type of technical assistance requested by participants

Type of request	**Scrolling – Number of requests (% of 19 requests)**	**Paging – Number of requests (% of 16 requests)**	**Row total (% of 35 requests across both groups)**
Clarification of question	14 (73.68%)	11 (68.75%)	25 (71.43%)
Interface issues	1 (5.26%)	1 (6.25%)	2 (5.71%)
Navigation issues	0 (0%)	3 (18.75%)	3 (8.57%)
Not recorded	4 (21.05%)	1 (6.25%)	5 (14.29%)

3.5.3. Smartphone and tablet computer ownership

From the 530 participants included in the analysis, 518 (97.74%) reported owning a smartphone and 435 (82.10%) reported owning a tablet PC.

3.6. Discussion

This study assessed the feasibility of using tablet computers in the waiting area of antenatal clinics for the implementation of NICE recommendations for the recognition of antenatal depression. Overall, approximately 23% of participants answered *Yes* to at least one Whooley question, and approximately 13% of participants scored 10 points or more on the EPDS. These are screening scales and, as such, do not provide a diagnosis of depression. However, these figures are within the expected prevalence rates of antenatal depression that have been reported previously [7-9]. This finding supports the idea that screening through a tablet computer app will not result in a disproportionate increase in the number of pregnant women screening positive for depression, which could add additional pressure on antenatal services. Thus, providing a certain degree of reassurance over concerns that self-administration and the perceived sense of anonymity created by mobile technology could encourage the over-reporting of sensitive information such as mental health symptoms [34].

Two survey layouts (i.e., Scrolling vs Paging) were compared in relation to participants' responses to Whooley questions and the EPDS. There was no relationship between survey layout (i.e., Scrolling and Paging) and (i) the proportion of participants screening positive (or negative) with the Whooley questions, (ii) participants' median EPDS scores, and (iii) participants' answers to question 10 of the EPDS (which deals with thoughts of self-harm).

However, there was a significant relationship between survey layout and EPDS scoring thresholds. This finding suggests a survey layout effect that manifests itself in pregnant women at the highest scoring interval of the EPDS. The implication for antenatal services is that the administration of the EPDS using a *Paging* layout could increase the risk of missing pregnant women who might have already met diagnostic criteria for major depression and require specialist support. Alternatively, this finding could indicate that a *Scrolling* layout might result in an increased number of false positives on the EPDS, leading to the unnecessary administration of further mental health assessments to these pregnant women. These explanations are compatible with the finding that the observed frequency of participants scoring 13 points or more on the EPDS (3% in the Paging group and 8% in the Scrolling group) differed significantly from the expected frequency for each experimental group. Although this argument has implications for the workload of antenatal care providers, it does not contradict

the finding that screening through mobile devices does not result in a disproportionate increase in the number of women screening positive. Instead, it relates to the unintended consequences of a survey layout on the effectiveness of screening activities and on the quality of care.

Diagnostic procedures were not part of this study and it was not possible to confirm if those participants in the 13+ EPDS scoring interval were indeed clinically depressed. Therefore, future exploration of this potential survey layout effect should be conducted using a similar study design where participants undergo diagnostic procedures (e.g., structured clinical interviews) after completing the screening scales. Such a study is necessary not only for replicating this effect, but also for determining its direction (i.e., identifying if it is due to an increase in the number of false positive or false negatives).

A future study should also consider the distribution of known risk factors for antenatal depression in the sample of participants, such as low social support, major life events, and history of abuse amongst others [35]. This is to corroborate that the survey layout effect is not driven by differences in these risk factors.

Concerning the other study outcomes, completion time and participants' requests for technical assistance were chosen as indicators of the impact that this approach to depression screening had on the workflows and schedules of antenatal clinics.

In relation to completion time, most participants (75%) completed the study procedures in 7 minutes 16 seconds or less. However, this measure included the socio-demographic survey, which is not a component of depression screening. As such, it represents an overestimation of the real duration of the screening process (i.e., Whooley questions and EPDS). These completion times were considered in relation to internal trials of the Whooley questions and the EPDS (which yielded an estimate of 2 minutes 28 seconds). This suggests that this form of antenatal depression screening might cause minimal disruption to a clinic's schedules, given that the time needed to complete the screening procedures did not appear to exceed patient waiting time (as suggested by the low breakoff rates, discussed further below). Participants in the scrolling group completed the surveys significantly faster than participants in the paging

group. However, this difference of 47.5 seconds might not be meaningful for clinical practice given the speed of completion.

Approximately 7% of all participants requested technical assistance during survey completion, which amounted to 35 unique requests (as no participant made more than one request). The choice of survey layout did not affect the proportion of participants asking for help in each group. About 71% of these requests (25 of 35) were to clarify the meaning of questions. The remaining 10 requests were related to difficulties navigating through the surveys (due to the implementation of validation rules that prevented participants from progressing through the surveys leaving questions unanswered) and interacting with the user interface (namely, interacting with the response options). These findings suggest that this population is already familiar with mobile technology, also implied by the level of smartphone ownership (approximately 98% of participants) and tablet computer ownership (approximately 82% of participants) in this sample of participants. Consequently, there is no need for clinical members of staff to oversee the survey completion process. Instead, these requests could be dealt with by non-clinical support staff (as long as patients are able to discuss their screening results with their clinician during their antenatal appointment).

A perceived concern that has been reported in the literature is the location and privacy of screening [5,22]. This concern was also voiced by many clinicians at the various recruitment sites prior initiating the study. Recruitment and completion rates were chosen as indicators of whether participants felt comfortable undergoing depression screening in the waiting area of antenatal clinics. Most eligible mothers-to-be who were approached (approximately 90%) agreed to take part in the study, and approximately 91% of those who took part completed the study procedures. In those sites where the research midwives or CSOs recorded the reasons given by eligible mothers-to-be for refusing to take part in the study, they expressed a lack of interest in the study and not wanting to miss being called to their appointment. Moreover, the only reason that some participants could not complete the study procedures (i.e., breakoff) was that they were called in to their appointment. These findings suggest that mobile technology may offer sufficient privacy and anonymity to patients, so that they are able to complete depression screening scales in a clinic's waiting area before their antenatal

appointment. These findings resonate with those of Kingston and colleagues [36], who also found that electronic mental health screening is feasible and acceptable to pregnant women.

Breakoff rates (and the corresponding reasons) were initially intended as an indicator of participants' perceived intrusiveness of the screening process. Since participants suspended their participation only if they were called in to their appointment, this outcome became an indicator of the compatibility between clinical workflows and research procedures. Overall, 133 women were called in to their appointment at different stages of the study: 44 (33%) before being approached by a research midwife or CSO, 35 (26%) before they were able to complete the consent process, and 54 (41%) before they were able to complete the study procedures.

The main difficulty during the initial stage was researchers' lack of time or capacity. Most women called in before being approached about participation (37 of 44 or approximately 84%) were in primary care facilities with short waiting times between patient check in and appointments. This presented a challenge for research midwives and CSOs working in these facilities, as they struggled to recruit enough participants into the study. The remaining women (7 of 44 or approximately 16%) were in secondary care sites with a larger volume of patients and longer waiting times. In these facilities however, research midwives and CSOs did not have the capacity to approach all the eligible mothers-to-be who were identified.

Although most women who were called in during the consent process or the study procedures were in secondary care facilities (73 of 89 – 82%), the main difficulty during these stages was the incompatibility between the administrative research procedures and clinical workflows. Firstly, the socio-demographic survey (which clinical care teams would not need to administer as they should have access to this information) inflated the real duration of the screening process. However, it is also important to consider the additional time required to conduct non-routine, non-clinical procedures and their impact on clinical workflows: providing information about the study, discussing a PIL, answering any questions that potential participants might have, and completing an informed consent form.

These findings highlight the tension between the need for sound evidence about the impact of technology on antenatal depression screening, and the challenges associated with

incorporating new technologies in real clinical settings. They also highlight that the planning of antenatal depression screening through mobile devices must be sensitive to local clinical workflows, and to the role of the staff members (both clinical and non-clinical) who will be responsible for their implementation.

The introduction of mobile technology for antenatal depression screening in clinical settings also needs to take into consideration issues of connectivity. Wi-Fi access varied between the participating clinics. This was a key determinant for the choice of app, as support for offline data collection was required. This might also represent a requirement for clinical teams wishing to implement this type of depression screening. In relation to connectivity, there is an added issue concerning the timely transmission of patient results to clinicians. In this study, this was done through a paper form that research midwives and CSOs filled in as soon as participants completed the study. This would introduce an extra administrative step to the routine activities of a clinic, not taking full advantage of the connectivity capabilities of mobile devices to facilitate data management. Timely transmission of patient results could be a problem without interoperability between a mobile app and clinical systems. Indeed, this has been identified as one of the barriers preventing the incorporation of patient data gathered through apps into existing workflows [37].

Lastly, the introduction of mobile technology in clinical settings also needs to take into account the management of technology. Security is a key concern in this context, particularly in relation to lost devices or theft. In addition to the economic implications, this could also lead to a breach of confidentiality, if patient data are held on the devices. Therefore, clinical teams should ensure compliance with their organisation's information governance policies without affecting patient experience. For example, passwords are considered ideal for protecting sensitive data. However, passwords could also disrupt clinical workflows if patients using password-protected devices time out and have to request help from staff members. Clinical teams also need to ensure that data are stored and transmitted using approved data encryption standards. Moreover, those responsible for the devices must be aware of any software update containing security patches.

3.6.1. Limitations

Diagnostic procedures were not conducted as part of this study. Therefore, it was not possible to confirm if the effect of survey layout on the proportion of pregnant women at each scoring interval of the EPDS was associated with cases of clinical depression.

Only adult participants who felt comfortable reading and writing in English were recruited in this study, which limits the generalisability of findings. Since it has been suggested that migrants might be at increased risk for mental health problems [38], the findings from this study might only be applicable to women at lower risk for antenatal depression. Particularly since women who were experiencing a common mental health disorder at the time of the study were also excluded. However, relying on self-disclosure when assessing the eligibility of potential participants made it impossible to confirm that these women were actually excluded from the study. In addition, participants were not assessed for the presence of other known risk factors for antenatal depression such as low social support, history of miscarriages or termination of pregnancy, history of abuse or the presence of other common mental health problems [35]. Future studies should aim to recruit a more representative sample of participants and explore potential differences in effect size of direction of effect.

Measures of completion time included the socio-demographic baseline survey, which is not a component of antenatal depression screening. Therefore, the inclusion of this survey inflated the real duration of the screening process. In some sites, the study procedures were administered by research staff with no links to the clinical team, or by research midwives with dual responsibilities: clinical and research. In addition, the administrative research requirements were unnatural to the reality of antenatal clinics and, in some cases, caused disruption. Future assessments of this approach to antenatal depression screening should evaluate implementation strategies that are compatible with the realities and complexities of antenatal services.

Moreover, indirect measures were used to assess the impact of mobile technology on the workflows and schedules of antenatal clinics, and the acceptability to patients. In addition, there was no qualitative assessment of the perceived usefulness to healthcare providers and patients of this form of depression screening, including how the results letters were used

during the antenatal appointments and how they influenced patients' pathways, if at all. These factors should be explored in future studies, with a particular focus on differences associated with the socio-demographic or clinical characteristics of the participants, as well as with local or regional arrangements for antenatal and mental health services.

3.6.2. Conclusions

This study suggests that it is feasible to use tablet computer apps for the administration of validated surveys in clinical settings: in this case, for the administration of depression screening scales in the waiting area of antenatal clinics. Moreover, the findings suggest that a scrolling or paging layout does not influence the proportion of pregnant women who screen positive with the Whooley questions. However, there was a significant relationship between survey layout and EPDS scoring intervals, which could have implications for the provision of antenatal mental health care: risk of missing pregnant women who might have already met the diagnostic criteria for depression, or the unnecessary administration of further mental health assessments to women who are not at increased risk of antenatal depression. This association should be explored further in relation to confirmed cases of clinical depression.

This study also evaluated the potential impact of mobile technology on the workload and timings of antenatal clinics. Tablet computer apps for antenatal depression screening do not seem to disrupt the routine activities of clinical staff, as measured by completion times and the proportion of participants asking for technical assistance. Nonetheless, it is crucial that practitioners and commissioners consider the successful integration of mobile technology into existing clinical workflows, the timing of appointment as well as connectivity and security issues.

The rates of smartphone and tablet computer ownership in this sample of participants, suggests that this population are already using mobile technology for personal reasons. This familiarity might thus enable them to interact with this technology in clinical settings without the need for supervision. The observed response and completion rates also suggest that this form of depression screening might be acceptable to this population.

References (Chapter 3)

1. BinDhim NF, Alanazi EM, Aljadhey H, Basyouni MH, Kowalski SR, Pont LG, Shaman AM, Trevena L, Alhawassi TM. Does a Mobile Phone Depression-Screening App Motivate Mobile Phone Users With High Depressive Symptoms to Seek a Health Care Professional's Help? *Journal of Medical Internet Research*. 2016;18(6): e156. Available from: doi: 10.2196/jmir.5726 [Accessed 01 March 2017].
2. BinDhim NF, Shaman AM, Trevena L, Basyouni MH, Pont LG, Alhawassi TM. Depression screening via a smartphone app: cross-country user characteristics and feasibility. *Journal of the American Medical Informatics Association (JAMIA).* 2015;22(1): 29-34. Available from: doi: 10.1136/amiajnl-2014-002840 [Accessed 01 March 2017].
3. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker Mobile Web App With Email for Adults to Self-Manage Mild-to-Moderate Depression: Randomized Controlled Trial. *JMIR Mhealth and Uhealth*. 2016;4(1): e8. Available from: doi: 10.2196/mhealth.4231 [Acessed 01 March 2017].
4. Faurholt-Jepsen M, Frost M, Vinberg M, Christensen EM, Bardram JE, Kessing LV. Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry Research*. 2014;217(1-2): 124-7. Available from: doi: 10.1016/j.psychres.2014.03.009 [Accessed 01 March 2017].
5. Tabb KM, Choi S, Pineros-Leano M, Meline B, McDonald HG, Kester R, Huang H. Perinatal depression screening in a Women, Infants, and Children (WIC) program: perception of feasibility and acceptability among a multidisciplinary staff. *Gen Hospital Psychiatry*. 2015;37(4): 305-9. Available from: doi: 10.1016/j.genhosppsych.2015.03.008 [Accessed 01 March 2017].
6. Marcano Belisario JS, Jamsek J, Huckvale K, O'Donoghue J, Morrison C, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. Cochrane Database of Systematic Reviews. 2015;7: MR000042. Available from: doi: 10.1002/14651858.MR000042.pub2 [Accessed on 01 March 2017].
7. Gavin NI, Gaynes BN, Lohr KN, Meltzer-Brody S, Gartlehner G, Swinson T. Perinatal depression: a systematic review of prevalence and incidence. *Obstetrics and Gynecology*. 2005; 106(5 Pt 1): 1071-83.

8. Bennett HA, Einarson A, Taddio A, Koren G, Einarson TG. Prevalence of depression during pregnancy: systematic review. *Obstetrics and Gynecology*. 2004;103(4): 698-709.
9. Heron J, O'Connor TG, Evans J, Golding J, Glover V, ALSPAC Study Team. The course of anxiety and depression through pregnancy and the postpartum in a community sample. *Journal of Affective Disorders*. 2004;80(1): 65-73.
10. Pearson RM, Evans J, Kounali D, Lewis G, Heron J, Ramchandani PG, O'Connor TG, Stein A. Maternal depression during pregnancy and the postnatal period: risks and possible mechanisms for offspring depression at age 18 years. *JAMA Psychiatry*. 2013;70(12): 1312-9. Available from: doi: 10.1001/jamapsychiatry.2013.2163 [Accessed 01 March 2017].
11. Kinsella MT, Monk C. Impact of maternal stress, depression and anxiety on fetal neurobehavioral development. *Clinical Obstetrics and Gynecology*. 2009;52(3): 425-40. Available from: doi: 10.1097/GRF.0b013e3181b52df1 [Accessed 01 March 2017].
12. Stewart DE. Clinical practice. Depression during pregnancy. *New England Journal of Medicine*. 2011;365(17): 1605-11. Available from: doi: 10.1056/NEJMcp1102730 [Accessed 01 March 2017].
13. Braithwaite EC, Murphy SE, Ramchandani PG. Effects of prenatal depressive symptoms on maternal and infant cortisol reactivity. *Archives of Women's Mental Health*. 2016; 19(4): 581-90. Available from: doi: 10.1007/s00737-016-0611-y [Accessed 01 March 2017].
14. Bauer A, Parsonage M, Knapp M, Iemmi V, Adelaja B. The cost of perinatal mental health problems. London, UK: Centre for Mental Health and London School of Economics; 2014.
15. Woolhouse H, Gartland D, Mensah F, Brown SJ. Maternal depression from early pregnancy to 4 years postpartum in a prospective pregnancy cohort study: implications for primary health care. *BJOG*. 2015;122(3):312-21. Available from: doi: 10.1111/1471-0528.12837 [Accessed 01 March 2017].
16. Howard LM, Molyneaux E, Dennis CL, Rochat T, Stein A, Milgrom J. Non-psychotic mental disorders in the perinatal period. *Lancet*. 2014;384(9956): 1775-88. Available from: doi: 10.1016/S0140-6736(14)61276-9 [Accessed 01 March 2017].

17. Kingston D, Austin MP, Heaman M, McDonald S, Lasiuk G, Sword W, Giallo R, Hegadoren K, Vermeyden L, van Zanten SV, Kingston J, Jarema K, Biringer A. Barriers and facilitators of mental health screening in pregnancy. *Journal of Affective Disorders*. 2015;186: 350-7. Available from: doi: 10.1016/j.jad.2015.06.029 [Accessed 01 March 2017].
18. Goodman JH. Women's attitudes, preferences, and perceived barriers to treatment for perinatal depression. *Birth*. 2009;36(1): 60-69. Available from: doi: 10.1111/j.1523-536X.2008.00296.x [Accessed 01 March 2017].
19. National Institute for Health and Clinical Excellence (2008 – 2017 Update). *Antenatal care for uncomplicated pregnancies*. Clinical Guideline CG62.
20. Sherwin H, McKeown M, Evans MF, Battacharyya OK. The waiting room "wait": from annoyance to opportunity. *Canadian Family Physician*. 2013;59(5): 479-481.
21. Mavletova A, Couper MP. Mobile web survey design: scrolling versus paging, SMS versus e-mail invitations. *Journal of Survey Statistics and Methodology*. 2014;2(4): 498-518. Available from: doi: 10.1093/jssam/smu015 [Accessed 01 March 2017].
22. Wells T, Bailey JT, Link MW. Comparison of Smartphone and Online Computer Survey Administration. *Social Science Computer Review*. 2014;32(2): 238-255. Available from: doi: 10.1177/0894439313505829 [Accessed 01 March 2017].
23. Pineros-Leano M, Tabb KM, Sears H, Meline B, Huang H. Clinic staff attitudes towards the use of mHealth technology to conduct perinatal depression screenings: a qualitative study. *Family Practice*. 2015;32(2): 211-5. Available from: doi: 10.1093/fampra/cmu083 [Accessed 01 March 2017].
24. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Morrison C, Car J. Tablet computers for implementing NICE antenatal mental health guidelines: protocol of a feasibility study. *BMJ Open*. 2016;6(1): e009930. Available from: doi: 10.1136/bmjopen-2015-009930 [Accessed 01 March 2017].
25. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Ramchandani P, Morrison C, Car J. Implementation of depression screening in antenatal clinics through tablet computers: results of a feasibility study. *BMC Medical Informatics and Decision Making*. 2017;17:59. Available from: doi: 10.1186/s12911-017-0459-8 [Accessed 11 May 2017].

26. National Collaborating Centre for Mental Health (2014 – 2015 Update), *Antenatal and postnatal mental health: clinical management and service guidance*. Clinical Guideline CG192.
27. Snap Surveys Limited. *Survey Software You'll Never Outgrow*. Available from: http://www.snapsurveys.com/survey-software/ [Accessed 01 March 2017].
28. Apple Inc. *iTunes Preview: Snap Mobile by Snap Surveys Limited.* Available from: https://itunes.apple.com/gb/app/snap-mobile/id507440809?mt=8 [Accessed 01 March 2017].
29. Snap Surveys Limited. *Snap WebHost – Online and Mobile Survey Management*. Available from: http://www.snapsurveys.com/survey-software/snap-webhost-online-mobile-survey-management-system/ [Accessed 01 March 2017].
30. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression – two questions are as good as many. *Journal of General Internal Medicine.* 1997;12:439-45.
31. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *British Journal of Psychiatry.* 1987; 150:782-786.
32. StataCorp. *Stata Statistical Software: Release 13* [Software]. College Station, Texas: StataCorp LP; 2013.
33. R Core Team. *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing, Vienna, Austria; 2016. Available from: https://www.R-project.org/.
34. Murdoch M, Simon AB, Polusny MA, Bangerter AK, Grill JP, Noorbaloochi S, Partin MR. Impact of different privacy conditions and incentives on survey response rate, participants representativeness, and disclosure of sensitive information: a randomized controlled trial. *BMC Medical Research Methodology*. 2014;14:90.
35. Leigh B, Milgrom J. Risk factors for antenatal depression, postnatal depression and parenting stress. *BMC Psychiatry*. 2008;8:24. Available from: doi: 10.1186/1471-244X-8-24 [Accessed 11 May 2017].
36. Kingston D, Austin MP, Veldhuyzen van Zanten S, Harvalik P, Giallo R, McDonald SD, MacQueen G, Vermeyden L, Lasiuk G, Sword W, Biringer A. Pregnant Women's Views on the Feasibility and Acceptability of Web-Based Mental Health E-Screening Versus

Paper-Based Screening: A Randomized Controlled Trial. *Journal of Medical Internet Research*. 2017;19(4):e88. Available from: doi: 10.2196/jmir.6866 [Accessed 11 May 2017].

37. West P, Van Kleek M, Giordano R, Weal MJ, Shadbolt N. Common Barriers to the Use of Patient-Generated Data Across Clinical Settings. *Proceedins of the 2018 Conference on Human Factors in Computing Systems (CHI),* 21 – 26 April 2018, Motreal, Canada. Available from: doi: 10.1145/3173574.3174058 [Accessed 09 August 2018].
38. Gilliver SC, Sundquist J, Li X, Sundquist K. Recent research in the mental health of immigrants to Sweden: a literature review. *European Journal of Public Health*. 2014;24(Suppl 1):72-79. Available from: doi: 10.1093/eurpub/cku101 [Accessed 09 August 2018].

4. Tablet computers for the assessment of the diagnostic performance of validated screening scales: the case of the Whooley questions

4.1. Overview

The diagnostic performance of depression screening scales can vary depending on the setting and the clinical population to which they are administered. Consequently, their endorsement for use in particular situations is sometimes based on limited evidence. This is the case of the Whooley questions, which constitutes the cornerstone of depression screening in antenatal populations in the NHS.

Mobile technology can play a key supporting role for the assessment of the diagnostic performance of validated screening scales. Mobile devices can enable the administration of these scales to a potentially large number of participants in a relatively short time span. This would be particularly beneficial in situations where the gold standard is also a self-administered survey. Nonetheless, even if the gold standard were a diagnostic interview, consumer mobile devices could facilitate the data management requirements during the assessment process.

This chapter describes the evaluation of the diagnostic performance of the Whooley questions in relation to EPDS scores. The high sensitivity and specificity of the Whooley questions observed in this study suggest that the current scoring approach of this scale proposed by NICE (i.e., an affirmative answer to either question) might be appropriate to identify mothers-to-be with EPDS scores above the threshold for probable depression in the general population. However, to determine the effectiveness of the two-stage approach to depression screening recommended by NICE, the diagnostic performance of the Whooley questions and the EPDS must be assessed against diagnostic procedures.

4.2. Rationale

Mobile technology could address some of the practical barriers preventing the timely identification of depression during pregnancy. The previous chapter explored two implementation challenges associated with the introduction of mobile technology in antenatal clinics (i.e., survey layout and deployment). However, the cost-effectiveness of screening practices has been cited as one of the barriers to antenatal depression screening. Particularly, in relation to the diagnostic performance of screening scales. This chapter explores how mobile technology could assist in evaluating, and gathering evidence to support, the screening scales recommended by NICE guidelines.

4.3. Background

As mentioned in previous chapters, depression is one of the most common mental health problems during pregnancy with potentially long-lasting consequences for mothers, their children and families, and society in general [1-5]. Depending on the trimester, its prevalence can vary between 7% and 13%, and approximately 12% of all pregnant women could experience an episode of major depression at any stage during their pregnancies [3, 6, 7]. Pregnant women struggling with depression are more likely to engage in unhealthy behaviours, and are at increased risk of postnatal depression and self-harm [1, 8, 9]. In addition, antenatal depression operates as an independent risk factor during a child's behavioural, cognitive and emotional development [1, 10-12].

Although there is no difference in the incidence and prevalence of depression between pregnant and non-pregnant women, the evidence suggests that rates of recognition and treatment are low during pregnancy [13, 14]. This has highlighted the need for improved identification strategies (combined with appropriate treatment programmes). However, it is unclear if depression screening during pregnancy is clinically and cost effective, and there is uncertainty about the most suitable screening instruments [14]. The performance of these instruments can vary depending on the setting, target population, language and scoring criteria [15]. In some cases, their use is recommended on the basis of limited evidence.

In the UK, NICE recommends using the Whooley questions to initiate the process of identifying depression during the perinatal period [16]. These two questions assess the presence of depressed mood or anhedonia during the past month [17]. An affirmative answer to either question should prompt further assessment with a validated screening tool such as the EPDS or the PHQ-9, the results of which will determine the need for further evaluation by a specialist [16]. Although the Whooley questions are a key element in recognising depression in perinatal populations, their endorsement is based on limited evidence of their diagnostic performance [9, 18, 19].

The diagnostic performance of a validated scale typically refers to its sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and false omission rate (FOR) calculated against a gold standard. This represents an important aspect of data quality, as the assessments made with these scales could trigger pre-specified diagnostic and referral pathways for patients. This aspect of data quality is predominantly linked to the characteristics of a scale. However, mobile technology could play a supporting role in generating evidence of the diagnostic performance of a given scale, in this case the Whooley questions.

The administration of the Whooley question, or any other scale, through mobile devices in a clinical setting offer the potential to reach a large sample of patients. This could be particularly advantageous for the evaluation of diagnostic performance if the gold standard is also a self-administered survey. The work discussed in this chapter was conducted as part of a study to assess the feasibility of using tablet computers in the waiting area of antenatal clinics for implementing NICE recommendations [16] for the recognition of antenatal depression (discussed in Chapter 3). Its aim was to assess the feasibility of using tablet computers to evaluate the diagnostic performance of the Whooley questions in relation to EPDS scores. The EPDS is not a diagnostic instrument and, as such, does not represent a gold standard. However, it was chosen as the gold standard for this study given the two-stage approach to depression screening recommended by NICE: administering the Whooley questions initially to discriminate those mothers-to-be who need to complete the EPDS; and using the latter to determine those who might need further specialist assessment.

4.4. Methods

This study received favourable opinion from the South East Coast – Surrey National Research Ethics Committee on 09 July 2015, under the reference number 15/LO/0977. The methodology described in this chapter was published as a study protocol in a peer reviewed journal [20].

4.4.1. Aim

This study was designed to evaluate the feasibility of using tablet computers to administer a bespoke socio-demographic survey, the Whooley questions, and the EPDS to 530 pregnant women who attended NHS antenatal clinics across England between October 2015 and May 2016. Chapter 3 presented findings concerning general feasibility, issues of survey layout, and deployment challenges. The aim of this chapter is to report findings concerning the diagnostic performance of the Whooley questions in relation to EPDS scores. The overall diagnostic performance of this scale was evaluated in relation to its sensitivity, specificity, PPV, NPV, and FOR. Furthermore, the diagnostic performance of each Whooley question was evaluated separately to determine if either of them was responsible for the overall performance of this screening instrument. In addition, the Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) of the Whooley Questions was calculated for different EPDS scoring thresholds [21, 22].

4.4.2. Participant recruitment and study procedures

Adult (18 years old or older) pregnant women of any gestational age, pregnancy history and parity were recruited from NHS antenatal clinics in England. Women who self-disclosed having a recent history (up to 12 months prior being approached by the research team) of depression or generalised anxiety disorder, who were currently receiving treatment for any of these disorders, or who did not feel comfortable reading and writing in English were not eligible to participate in this study.

Participants were recruited using an opportunistic approach. Potential participants were approached in the waiting area of antenatal clinics by a research midwife or a CSO with information about the study and a copy of the PIL. They were also given the opportunity to

ask any questions about the study before deciding on participation. Women consenting to take part in the study were then asked to complete and sign an informed consent form, after which they completed the study procedures. Participants were required to complete their participation before their appointment. If a participant was called in before submitting her responses, her participation was suspended and her data deleted, and she was not re-approached afterwards.

Recruitment took place in 14 localities across England: Blackburn, Bolton, Burton, Bury, East Midlands, London, Huntingdon, North East, Oldham, Rochdale, Shrewsbury, Stoke-on-Trent, Telford and Wigan. Between October 2015 and May 2016, 743 pregnant women were identified in the waiting area of participating clinics. Of these, 699 were asked to take part in the study (as 44 were called in before a research midwife or CSO could approach them). Six hundred and thirty-two women agreed to take part and initiated the consent process, but only 597 were randomised into the study (as 35 were called in whilst completing the consent form). Finally, 543 women completed the study procedures, after 54 were called in to their appointment, of which 530 were included in the analysis (as 13 records had to be removed because they were submitted more than 24 hours after completion).

4.4.3. Surveys

4.4.3.1. Socio-demographic survey

A bespoke socio-demographic survey consisting of 11 questions was used to gather information about participants' age group, ethnic background, relationship status, employment status, highest educational achievement, smartphone and tablet ownership, obstetric history, pregnancy trimester, and previous history of depression (Appendix 32).

4.4.3.2. Depression screening scales

4.4.3.2.1. Whooley questions

As per NICE recommendations [16], participants were asked to complete the Whooley questions. This is a 2-item survey assessing the presence of depressed mood and anhedonia over the past month [17] (Appendix 33). An affirmative answer to either question should prompt further assessment using a validated screening instrument such as the EPDS.

4.4.3.2.2. Edinburgh Postnatal Depression Scale

Participants were asked to complete the EPDS. This is a 10-item depression screening scale that has been validated for use in the antenatal or the postnatal period, as well as in the community or in clinical settings [15]. This scale assesses feelings of guilt, sleep disturbance, anhedonia and suicidal ideation that have been present for the past 7 days (Appendix 34). Each question is scored on a 4-point scale from 0 to 3 points. Individual responses are then added together to generate an overall score, which is indicative of a respondent's risk of depression. The performance of the EPDS can vary according to the setting in which it is administered and the sample of patients. However, the following cut-off points have been reported in the literature: 9 or 10 points for possible depression, 12 or 13 points for probable depression, and 15 points for antenatal depression [15]. In addition, scores of 1 point or more on question 10 of this scale (regardless of the overall score) is suggestive of increased risk of self-harm. In any of these scenarios, patients should undergo formal diagnostic assessment.

NICE guidelines [16] recommend that an affirmative answer to either Whooley question should prompt further assessment using a validated instrument such as the EPDS. However, in this study all participants were asked to complete the EPDS regardless of their answers to the Whooley questions.

4.4.4. Technology and materials

The surveys were implemented using Snap® survey software [23], and administered through Snap® Mobile app (versions 4.0.30 to 4.0.32) for Apple® iOS™ [24] running on Apple® iPad Air® tablet computers. Responses were stored using Snap® WebHost services [25].

4.4.5. Statistical analyses

The following outcomes were reported: participants' overall responses to the Whooley questions; participants' mean overall EPDS scores (and SD); and the proportion of participants at each scoring interval of the EPDS (i.e., 0 to 9 points indicating low risk of depression, 10 to 12 points suggestive of possible depression, 13 or more points suggestive of probable depression, and 15 points or more as indicative of probable depression in an antenatal sample

[15]). For the latter outcome, a breakdown by pregnancy trimester, obstetric history, ethnic background (white versus black ethnic minority groups), and previous history of depression was also given. In addition, the proportion of participants at increased risk of self-harm, as indicated by question 10 of the EPDS, was also presented.

In Chapter 3, 13 points on the EPDS was used as the threshold for probable depression, as this was in agreement with local care pathways in the participating clinics. Since this chapter is concerned with the evaluation of the diagnostic performance of the Whooley questions, all EPDS scoring thresholds were considered. Therefore, the diagnostic performance of the Whooley questions was assessed in terms of their specificity, sensitivity, PPV, NPV, and FOR in relation to two EPDS thresholds: 13 and 15 points. Furthermore, the diagnostic performance of each Whooley question was calculated separately. In addition, the AUC of the Whooley Questions was calculated for different EPDS scoring thresholds. For this analysis, a decision was made to explore only those scores for which the AUC was 0.8 or above, as this indicates a test with "very good" discriminative power [22].

4.4.5.1. Sensitivity

Overall sensitivity was measured as the number of participants who answered *Yes* to any Whooley question AND had an overall EPDS score of 13 points or more, as a proportion of the total number of participants with an overall EPDS score of 13 points or more (regardless of their answers to the Whooley questions). This outcome was also calculated using an EPDS threshold of 15 points. In addition, this outcome was measured for each Whooley question separately.

4.4.5.2. Specificity

Specificity was measured as the number of participants who answered *No* to both Whooley questions AND had an overall EPDS score of 12 points or less, as a proportion of the total number of participants with an overall EPDS score of 12 points or less (regardless of their answers to the Whooley questions). This outcome was also calculated using an EPDS threshold of 14 points or less. Moreover, specificity was calculated for each Whooley question separately.

4.4.5.3. Positive Predictive Value

PPV was measured as the number of participants who answered *Yes* to any Whooley question AND had an overall EPDS score of 13 points or more, as a proportion of the total number of participants who answered *Yes* to any Whooley question (regardless of their overall EPDS score). This outcome was also calculated using an EPDS threshold of 15 points. Moreover, PPV was calculated for each Whooley question separately.

4.4.5.4. Negative Predictive Value

NPV was measured as the number of participants who answered *No* to both Whooley questions AND had an overall EPDS score of 12 points or less, as a proportion of the total number of participants who answered *No* to both Whooley questions (regardless of their overall EPDS score). This outcome was also calculated using an EPDS threshold of 14 points or less. Moreover, NPV was calculated for each Whooley question separately.

4.4.5.5. False Omission Rate

FOR was measured as the number of participants who answered *No* to both Whooley questions AND had an overall EPDS score of 13 points or more, as a proportion of the total number of participants who answered *No* to both Whooley questions (regardless of their overall EPDS score). This outcome was also calculated using an EPDS threshold of 15 points. Moreover, FOR was calculated for each Whooley question separately.

For all outcomes, the 95% CI were calculated using the continuity corrected Wilson's procedure.

4.4.5.6. Area Under the Receiver Operating Characteristics (ROC) Curve (AUC)

ROC curves (and the corresponding AUC) were plotted for the Whooley Questions at EPDS thresholds ranging from 10 to 18 points, as these resulted in AUC values of 0.8 or higher. However, only the AUC for 12, 13 and 14 points were compared as these were the only EPDS scoring thresholds for which the lower 95% CI limit did not include values under 0.8. The aim of this approach was to compare the ability of the Whooley questions to discriminate

between pregnant women scoring at or above these scoring thresholds, and those scoring below.

All statistical analyses were conducted in R version 3.3.1 [26].

4.5. Results

4.5.1. Characteristics of the participants

Table 10 provides a summary of the socio-demographic characteristics of the participants.

Table 10. Participants' socio-demographic characteristics

		n (row %, N = 530)
Recruitment Setting	***General Practice***	115 (21.70%)
	Community	26 (4.91%)
	Hospital	389 (73.40%)
Age group	***18 – 22 years***	26 (4.91%)
	23 – 27 years	85 (16.03%)
	28 – 32 years	198 (37.36%)
	33 – 37 years	164 (30.94%)
	38+ years	57 (10.75%)
Ethnic Background	***White***	355 (66.98%)
	Mixed	16 (3.02%)
	Asian	109 (20.57%)
	Black	28 (5.28%)
	Other ethnic group	21 (3.96%)
	Not disclosed	1 (0.19%)
Relationship status	***Single***	30 (5.66%)
	Married/Civil Partnership	348 (65.66%)
	Cohabiting	138 (26.04%)
	Divorced/Widowed/Separated	5 (0.94%)
	Not disclosed	9 (1.70%)
Employment status	***Full-time employment***	276 (52.08%)
	Part-time employment	101 (19.06%)
	Self-employed	44 (8.30%)
	Seeking employment	19 (3.58%)
	Unemployed	83 (15.66%)
	Disability	7 (1.32%)
Highest educational achievement	***Postgraduate degree***	108 (20.38%)
	University degree	201 (37.92%)
	College degree or below	192 (36.23%)
	None	13 (2.45%)
	Other qualification	16 (3.02%)

4.5.2. Participants' responses to the Whooley questions and the Edinburgh Postnatal Depression Scale

Overall, 408 (76.98%) participants answered *No* to both Whooley questions, 72 (13.58%) participants answered *Yes* to one Whooley question (44 participants to question one and 28 participants to question 20), and 50 (9.43%) participants answered *Yes* to both Whooley questions. Moreover, the mean overall EPDS score was 4.58 points (SD = 4.35). In addition, 461 (86.98%) participants had overall EPDS scores between 0 and 9 points, and 38 (7.17%) participants had scores between 10 and 12 points. Thirty-one (5.85%) participants had scores of 13 points or more on the EPDS, whereas 21 (3.96%) participants had EPDS scores of 15 points or more on the EPDS (i.e., the threshold recommended for antenatal populations). Table 11 presents a summary of the proportion of participants at each EPDS scoring interval, as well as a breakdown by pregnancy trimester, obstetric history, ethnic background, and previous history of depression. In addition, 11 (2.08%) participants scored 1 point or more on question 10 of the EPDS, which deals with thoughts of self-harm.

Table 11. Number of participants at each EPDS scoring interval, and by pregnancy trimester, obstetric history, ethnic background and previous history of depression

		0-9 points n (row %)	10-12 points n (row %)	13+ points n (row %)	15+ points n (row %)[2]
Overall sample (N = 530)		461 (86.98)	38 (7.17)	31 (5.85)	21(3.96)
Pregnancy trimester	First (N = 79)	66 (83.54)	5 (6.33)	8 (10.13)	6 (7.59)
	Second (N = 191)	161 (84.29)	19 (9.95)	11 (5.76)	6 (3.14)
	Third (N = 260)	234 (90)	14 (5.38)	12 (4.62)	9 (3.46)
Obstetric history	Primigravida (N = 204)	175 (85.78)	16 (7.84)	13 (6.37)	9 (4.41)
	Multigravida (N = 326)	286 (87.73)	22 (6.75)	18 (5.52)	12 (3.68)
Ethnic background	White (N = 355)	318 (89.58)	22 (6.19)	15 (4.23)	10 (2.82)
	Black & minority ethnic (N = 175)	143 (81.71)	16 (9.14)	16 (9.14)	11 (6.29)
Previous history of depression	Yes (N = 59)	43 (72.88)	9 (15.25)	7 (11.86)	4 (6.78)
	No (N = 471)	418 (88.75)	29 (6.16)	24 (5.09)	17 (3.61)

4.5.3. Diagnostic performance of the Whooley questions

Table 12 shows the number of participants answering *Yes* to at least one Whooley question and those answering *No* to both questions against an EPDS scoring threshold of 13 points.

[2] Row percentages on this column reflect the change in the proportion of participants in relation to a cut-off score of 13 points

Table 12. Number of participants according to their answers to the Whooley questions and EPDS scores (EPDS threshold: 13 points)

	Yes to at least one Whooley question	No to both Whooley questions	Row total
13 points or more on the EPDS	28	3	**31**
Less than 13 points on the EPDS	94	405	**499**
Total	**122**	**408**	**530**

Table 13 presents the same information as Table 12 but using an EPDS scoring threshold of 15 points.

Table 13. Number of participants according to their answers to the Whooley questions and EPDS scores (EPDS threshold: 15 points)

	Yes to at least one Whooley question	No to both Whooley questions	Row total
15 points or more on the EPDS	18	3	**21**
Less than 15 points on the EPDS	104	405	**509**
Total	**122**	**408**	**530**

The information contained in Table 12 and Table 13 was used to assess the overall diagnostic performance of the Whooley questions in relation to the EPDS.

Tables 14 and Table 15 show the number of participants who answered *Yes* to question one of the Whooley questions against EPDS scoring thresholds of 13 and 15 points, respectively. Table 16 and Table 17 show the number of participants who answered *Yes* to question two of the Whooley questions against EPDS scoring thresholds of 13 and 15 points, respectively. Lastly, Table 18 and Table 19 show the number of participants who answered *Yes* to both Whooley questions against EPDS scoring thresholds of 13 and 15 points, respectively. The information provided by Tables 14 to 19 was used to calculated the diagnostic properties of each Whooley question separately.

Table 14. Number of participants according to their answers to question one of the Whooley questions and EPDS scores (EPDS threshold: 13 points)

	Yes to question one	No to question one	Row total
13 points or more on the EPDS	24	7	**31**
Less than 13 points on the EPDS	70	429	**499**
Total	**94**	**436**	**530**

Table 15. Number of participants according to their answers to question one of the Whooley questions and EPDS scores (EPDS threshold: 15 points)

	Yes to question one	No to question one	Row total
15 points or more on the EPDS	16	5	**21**
Less than 15 points on the EPDS	78	431	**509**
Total	**94**	**436**	**530**

Table 16. Number of participants according to their answers to question two of the Whooley questions and EPDS scores (EPDS threshold: 13 points)

	Yes to question two	No to question two	Row total
13 points or more on the EPDS	23	8	**31**
Less than 13 points on the EPDS	55	444	**499**
Total	**78**	**452**	**530**

Table 17. Number of participants according to their answers to question two of the Whooley questions and EPDS scores (EPDS threshold: 15 points)

	Yes to question two	No to question two	Row total
15 points or more on the EPDS	15	6	**21**
Less than 15 points on the EPDS	63	446	**509**
Total	**78**	**452**	**530**

Table 18. Number of participants who answered Yes to both Whooley questions according to their EPDS scores (EPDS threshold: 13 points)

	Yes to both questions	No to at least one question	Row total
13 points or more on the EPDS	19	12	**31**
Less than 13 points on the EPDS	31	468	**499**
Total	**50**	**480**	**530**

Table 19. Number of participants who answered Yes to both Whooley questions according to their EPDS scores (EPDS threshold: 15 points)

	Yes to both questions	No to at least one question	Row total
15 points or more on the EPDS	13	8	**21**
Less than 15 points on the EPDS	37	472	**509**
Total	**50**	**480**	**530**

4.5.3.1. Sensitivity

Of 31 participants with an EPDS score of 13 points or more, 28 answered *Yes* to at least one Whooley question (see Table 12). Therefore, the sensitivity of the Whooley questions using this scoring threshold was 90.32% (95% CI 75.10 to 96.65).

Of 21 participants with an EPDS score of 15 points or more, 18 answered *Yes* to at least one Whooley question (see Table 13). Therefore, the sensitivity of the Whooley questions using this scoring threshold was 85.71% (95% CI 65.36 to 95.02).

Of 31 participants with an EPDS score of 13 points or more, 24 answered *Yes* to question one of the Whooley questions (see Table 14), 23 answered *Yes* to question two of the Whooley questions (see Table 16) and 19 answered *Yes* to both Whooley questions (see Table 18). Therefore, the sensitivity of each question for this EPDS scoring threshold was 77.42% (95% CI 60.19 to 88.61), 74.19% (95% CI 56.75 to 86.30) and 61.29% (95% CI 43.82 to 76.27), respectively.

Lastly, of 21 participants with an EPDS score of 15 points or more, 16 answered *Yes* to question one of the Whooley questions (see Table 15), 15 answered *Yes* to question two of the Whooley questions (see Table 17) and 13 answered *Yes* to both Whooley questions (see Table 19). Therefore, the sensitivity of each question for this EPDS scoring threshold was 76.19% (95% CI 54.91 to 89.37), 71.43 (95% CI 50.05 to 86.19) and 61.90% (95% CI 40.87 to 79.25), respectively.

4.5.3.2. Specificity

Of 499 participants with an EPDS score of less than 13 points, 405 answered *No* to both Whooley questions (see Table 12). Therefore, the specificity of the Whooley questions using this scoring threshold was 81.16% (95% CI 77.50 to 84.35).

Of 509 participants with an EPDS score of less than 15 points, 405 answered *No* to both Whooley questions (see Table 13). Therefore, the specificity of the Whooley questions using this scoring threshold was 79.57% (95% CI 75.85 to 82.84).

Of 499 participants with an EPDS score of less than 13 points, 429 answered *No* to question one of the Whooley questions (see Table 14), 444 answered *No* to question two of the Whooley questions (see Table 16) and 468 answered *No* to at least one Whooley question (see Table 18). Therefore, the specificity of each Whooley question for this EPDS scoring threshold was 85.97% (95% CI 82.65 to 88.74), 88.98% (95% CI 85.93 to 91.44) and 93.79% (95% CI 91.32 to 95.59), respectively.

Of 509 participants with an EPDS score of less than 15 points, 431 answered *No* to question one of the Whooley questions (see Table 15), 446 answered *No* to question two of the Whooley questions (see Table 17) and 472 answered *No* to at least one Whooley question (see Table 19). Therefore, the specificity of each Whooley question for this EPDS scoring threshold was 84.68% (95% CI 81.29 to 87.55), 87.62% (95% CI 84.47 to 90.20) and 92.73% (95% CI 90.14 to 94.68), respectively.

4.5.3.3. Positive Predictive Value

Of 122 participants who answered *Yes* to at least one Whooley question, 28 had an EPDS score of 13 points or higher (see Table 12). Therefore, the PPV of the Whooley questions using this scoring threshold was 22.95% (95% CI 16.38 to 31.17).

Of the same 122 participants with an affirmative answer to at least one Whooley question, 18 had an EPDS score of 15 points or higher (see Table 13). Therefore, the PPV of the Whooley questions using this scoring threshold was 14.75% (95% CI 9.54 to 22.12).

Of 94 participants who answered *Yes* to question one of the Whooley questions, 24 had an EPDS score of 13 points or more (see Table 14). Therefore, the PPV of question one for this EPDS scoring threshold was 17.80% (95% CI 17.80 to 35.18). Of 78 participants who answered *Yes* to question two of the Whooley questions, 23 had an EPDS score of 13 points or more (see Table 16). Therefore, the PPV of question two for this EPDS scoring threshold was 29.49% (95% CI 20.53 to 40.38). Of 50 participants who answered *Yes* to both Whooley questions, 19 had an EPDS score of 13 points or more (see Table 18). Therefore, the PPV of both questions combined for this EPDS scoring threshold was 38.00% (95% CI 25.86 to 51.85).

With regard to the other EPDS scoring threshold, of 94 participants who answered *Yes* to question one of the Whooley questions, 16 had an EPDS score of 15 points or more (see Table 15). Therefore, the PPV of question one for this EPDS scoring threshold was 17.02% (95% CI 10.76 to 25.87). Of 78 participants who answered *Yes* to question two of the Whooley questions, 15 had an EPDS score of 15 points or more (see Table 17). Therefore, the PPV of question two for this EPDS scoring threshold was 19.23% (95% CI 12.01 to 29.33). Of 50 participants who answered *Yes* to both Whooley questions, 13 had an EPDS score of 15 points of more (see Table 19). Therefore, the PPV of both questions combined for this EPDS scoring threshold was 26.00% (95% CI 15.87 to 39.55).

4.5.3.4. Negative Predictive Value

Of 408 participants who answered *No* to both Whooley questions, 405 had an EPDS score of less than 13 points (see Table 12). Therefore, the NPV of the Whooley questions using this

scoring threshold was 99.26% (95% CI 97.86 to 99.75). The same result was obtained using 15 points as the scoring threshold (see Table 13).

Of 436 participants who answered *No* to question one of the Whooley questions, 429 had an EPDS score of less than 13 points (see Table 14). Therefore, the NPV of this question for this EPDS scoring threshold was 98.39% (95% CI 96.72 to 99.22). Of 452 participants who answered *No* to question two of the Whooley questions, 444 had an EPDS score of less than 13 points (see Table 16). Therefore, the NPV of this question for this EPDS scoring threshold was 98.23% (95% CI 96.55 to 99.10). Of 480 participants who answered *No* to at least one Whooley question, 468 had an EPDS score of less than 13 points (see Table 18). Therefore, the NPV of both questions combined for this EPDS scoring threshold was 97.50% (95% CI 95.68 to 98.56).

With regard to the other EPDS scoring threshold, of 436 participants who answered *No* to question one of the Whooley questions, 431 had an EPDS score of less than 15 points (see Table 15). Therefore, the NPV of this question for this EPDS scoring threshold was 98.85% (95% CI 97.34 to 99.51). Of 452 participants who answered *No* to question two of the Whooley questions, 446 had an EPDS score of less than 15 points (see Table 17). Therefore, the NPV of this question for this EPDS scoring threshold was 98.67% (95% CI 97.13 to 99.39). Of 480 participants who answered *No* to at least one Whooley question, 472 had an EPDS score of less than 15 points (see Table 19). Therefore, the NPV of both questions combined for this EPDS scoring threshold was 98.33% (95% CI 96.74 to 99.15).

4.5.3.5. False Omission Rate

Of 408 participants who answered *No* to both Whooley questions, three had an EPDS score of 13 points or more (see Table 12). Therefore, the FOR of the Whooley questions using this scoring threshold was 0.74% (95% CI 0.25 to 2.14). The same result was obtained when using 15 points as the scoring threshold (see Table 13).

Of 436 participants who answered *No* to question one of the Whooley questions, seven had an EPDS score of 13 points or more (see Table 14). Therefore, the FOR of this question for this EPDS scoring threshold was 1.61% (95% CI 0.78 to 3.28). Of 452 participants who answered

No to question two of the Whooley questions, eight had an EPDS score of 13 points or more (see Table 16). Therefore, the FOR of this question for this EPDS scoring threshold was 1.77% (95% CI 0.90 to 3.45). Of 480 participants who answered *No* to at least one Whooley question, 12 had an EPDS score of 13 points or more (see Table 18). Therefore, the FOR of both questions combined for this EPDS scoring threshold was 2.50% (95% CI 1.44 to 4.32).

With regard to the other EPDS scoring threshold, of 436 participants who answered *No* to question one of the Whooley questions, five had an EPDS score of 15 points or more (see Table 15). Therefore, the FOR of this question for this EPDS scoring threshold was 1.15% (95% CI 0.49 to 2.66). Of 452 participants who answered *No* to question two of the Whooley questions, six had an EPDS score of 15 points or more (see Table 17). Therefore, the FOR of this question for this EPDS scoring threshold was 1.33% (95% CI 0.61 to 2.87). Of 480 participants who answered *No* to at least one Whooley question, eight had an EPDS score of 15 points or more (see Table 19). Therefore, the FOR of both questions combined for this EPDS scoring threshold was 1.67% (95% CI 0.85 to 3.26).

Table 20 presents a summary of the diagnostic performance of the Whooley questions in relation to the two EPDS scoring thresholds assessed in this chapter.

Table 20. Summary of the diagnostic performance of the Whooley questions in relation to two EPDS scoring thresholds (i.e., 13 and 15 points)

		Overall diagnostic performance % (95% CI)	Question one of the Whooley Questions % (95% CI)	Question two of the Whooley Questions % (95% CI)	Both Whooley Questions Combined % (95% CI)
Sensitivity	EPDS: 13+ points	90.32 (75.10 to 96.65)	77.42 (60.19 to 88.61)	74.19 (56.75 to 86.30)	61.29 (43.82 to 76.27)
	EPDS: 15+ points	85.71 (65.36 to 95.02)	76.19 (54.91 to 89.37)	71.43 (50.05 to 86.19)	61.90 (40.87 to 79.25)
Specificity	EPDS: 13+ points	81.16 (77.50 to 84.35)	85.97 (82.65 to 88.74)	88.98 (85.93 to 91.44)	93.79 (91.32 to 95.59)
	EPDS: 15+ points	79.57 (75.85 to 82.84)	84.68 (81.29 to 87.55)	87.62 (84.47 to 90.20)	92.73 (90.14 to 94.68)
PPV	EPDS: 13+ points	22.95 (16.38 to 31.17)	17.80 (17.80 to 35.18)	29.49 (20.53 to 40.38)	38.00 (25.86 to 51.85)
	EPDS: 15+ points	14.75 (9.54 to 22.12)	17.02 (10.76 to 25.87)	19.23 (12.01 to 29.33)	26.00 (15.87 to 39.55)
NPV	EPDS: 13+ points	99.26 (97.86 to 99.75)	98.39 (96.72 to 99.22)	98.23 (96.55 to 99.10)	97.50 (95.68 to 98.56)
	EPDS: 15+ points		98.85 (97.34 to 99.51)	98.67 (97.13 to 99.39)	98.33 (96.74 to 99.15)
FOR	EPDS: 13+ points	0.74 (0.25 to 2.14	1.61 (0.78 to 3.28)	1.77 (0.90 to 3.45)	2.50 (1.44 to 4.32)
	EPDS: 15+ points		1.15 (0.49 to 2.66)	1.33 (0.61 to 2.87)	1.67 (0.85 to 3.26)

4.5.3.6. Area Under the ROC Curve (AUC)

Table 21 shows the AUC values and 95% CIs of the Whooley Questions at each of the EPDS thresholds for which the AUC exceeded 0.80. The highest AUC value was observed when 13 points on the EPDS was used as a threshold.

Table 21. AUC of the Whooley Questions for different EPDS thresholds

EPDS Threshold	AUC (95% CI)
10	0.8138 (0.7566 to 0.8711)
11	0.8569 (0.7978 to 0.916)
12	0.8757 (0.8175 to 0.9338)
13	0.8871 (0.8271 to 0.9471)
14	0.8793 (0.8085 to 0.9502)
15	0.8585 (0.7728 to 0.9442)
16	0.8254 (0.7093 to 0.9416)
17	0.8254 (0.7093 to 0.9416)
18	0.8223 (0.6777 to 0.9669)

Figure 10 shows the ROC curve of the Whooley Questions for those EPDS thresholds for which the lower 95% CI limit was above 0.8. These AUCs were compared using DeLong's test [22]. There was no statistically significant difference in the ability of the Whooley Questions to discriminate between pregnant women scoring at or above these EPDS thresholds, and pregnant women scoring below.

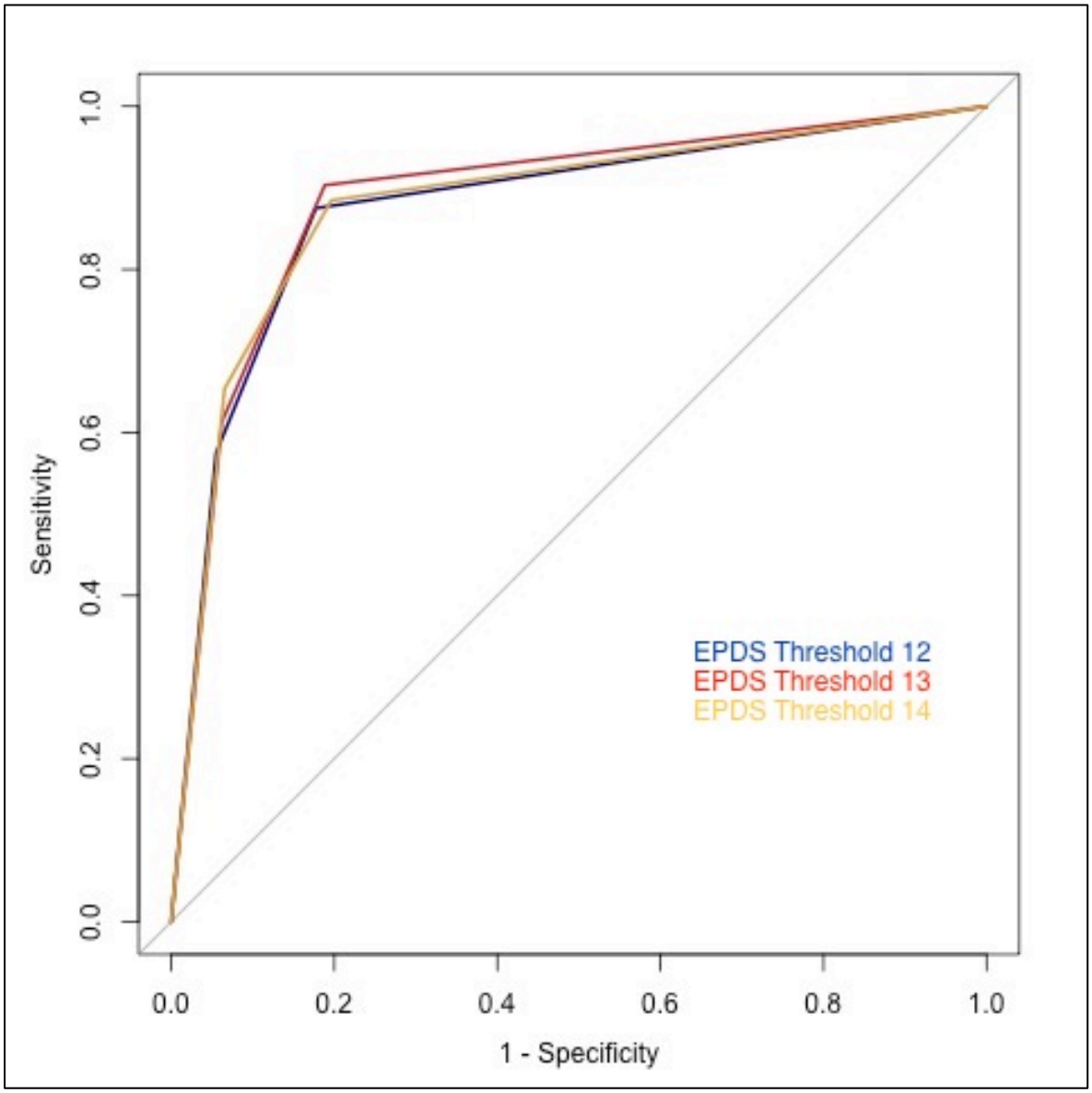


Figure 10. ROC curves of the Whooley Questions for the following EPDS thresholds: 12, 13 and 14 points

4.6. Discussion

This study evaluated the diagnostic properties (i.e., sensitivity, specificity, PPV, NPV and FOR) of the Whooley questions using a sample of participants recruited from an antenatal population. These properties were calculated in relation to two scoring thresholds of the EPDS: 13 points, and 15 points. The former is typically considered as an indicator of probable depression in the general population, whereas a score of 15 points is thought to indicate antenatal depression [15].

In terms of overall performance, the high sensitivity (90.32%) suggests that, in the context of antenatal clinics in the NHS, the recommended scoring procedure for the Whooley questions (i.e., an affirmative answer to at least one question) might be appropriate for recognising pregnant women with EPDS scores above the threshold for probable depression in the general population (i.e., 13 points). Moreover, the relatively high specificity (81.16%) suggests that the Whooley Questions can correctly identify a high proportion of pregnant women who would screen negative for antenatal depression. Increasing the EPDS scoring threshold to the one recommended for an antenatal population (i.e., 15 points) resulted in a decrease of the sensitivity, specificity and PPV of the Whooley questions by 5%, 2% and 8%, respectively. The NPV and FOR however, remained unchanged. Since the aim of depression screening is arguably not to miss those who might benefit from further assessment, it might be advisable to implement the EPDS scoring thresholds of probable depression for the general population in antenatal settings. Particularly since it was found that in some clinics, the EPDS scores are used to trigger formal mental health care pathways.

Furthermore, the PPV indicates that a low proportion of women screening positive with the Whooley questions had EPDS scores above the scoring threshold. However, the PPV is not an intrinsic property of the Whooley questions. Instead, it depends on the prevalence of antenatal depression, and on the sensitivity and specificity of the Whooley questions. The NPV and FOR however, indicate that there was a high probability that women screening negative with the Whooley questions had EPDS scores below the scoring thresholds.

Taken together, these results support the two-stage approach for the recognition of antenatal depression recommended by NICE. For example, if the second stage (i.e., the administration

of the EPDS) was removed, it could seriously affect the workload and workflows of antenatal clinics and associated mental health services. Although the Whooley questions have a high sensitivity, the wide 95% CI for this indicator and their low PPV suggest that further mental health assessments on the basis of at least one affirmative answer to either question would lead to a low proportion of pregnant women meeting the EPDS threshold currently used for triggering this pathway. This could then lead to an unnecessary increase in the workload of specialist mental health professionals due to false positives. When considering the negative answers given to the Whooley Questions, the high specificity and NPV and the low FOR suggest that the risk of missing women who might require further assessment (as indicated by the EPDS) might be low.

The diagnostic performance of each Whooley question and of both questions combined were evaluated separately. The aim was to determine if a different scoring approach could be more appropriate for identifying those pregnant women who require further mental health assessment using the EPDS. These analyses showed a decrease in the diagnostic performance of the Whooley Questions, except for specificity and PPV (for which question two performed slightly better than question one). Therefore, these findings suggest that the current scoring approach (i.e., at least one affirmative answer) in antenatal care is more appropriate.

Furthermore, the AUC (using ROC analysis) was calculated for the Whooley questions at various EPDS thresholds. Although the AUC values for EPDS thresholds between 10 and 18 points exceeded 0.80, only the 95% CIs for 12, 13 and 14 points were above this value. Moreover, no difference was observed in the ability of the Whooley Questions to discriminate between pregnant women scoring above and below these three EPDS thresholds. This further supports the recommendation of implementing the EPDS scoring thresholds of the general population in antenatal settings.

These findings resonate with those of Littlewood and colleagues [27]. These authors evaluated the diagnostic properties of the Whooley questions (and of the EPDS) against a diagnostic gold standard and concluded that although this instrument has acceptable sensitivity and specificity, its use in clinical practice is limited by its low predictive value and variable acceptability. They also found that using a single case-finding instrument (e.g. the

Whooley questions alone) was not the most cost-effective strategy for identifying perinatal depression. Instead, a two-stage case-finding strategy was more cost-effective [27].

Overall, this chapter highlighted important aspects about data quality and mobile technology. The diagnostic properties of a validated scale can be conceptualised as a data quality dimension. Since responses collected through these scales will inform patient care, they should accurately measure the underlying clinical construct. For this reason, their endorsement should be based on sound evidence of their performance. The previous chapter explored how the adaptation of the Whooley questions and the EPDS for use with a mobile app could affect the quality of the responses collected from an implementation perspective. This chapter on the other hand, explored how mobile technology can be harnessed to generate evidence of the diagnostic performance of a scale in a specific population (i.e., Whooley questions in antenatal populations).

Mobile phones can allow practitioners and policy makers to administer validated scales in large numbers. In cases where a diagnostic gold standard is also a self-administered survey, mobile technology could enable the collection of several data points in a relatively short time span. Even if the diagnostic standard is an interview or a lab test, mobile technology could facilitate the data management requirements of this process.

However, the real advantage of mobile technology is that it could, at least in theory, enable the validation of clinical scales in new settings and against new gold standards. Given the portability and usage patterns of smartphones and tablet computers, clinical scales could be easily administered outside clinical settings. Moreover, these devices could provide a more granular geographical notion of validation setting. For example, non-clinical settings could be further divided into home, office, commuting, shopping, and cinema amongst others. In addition, mobile phones could expand our geographical notion of validity by incorporating situational and contextual variables such as time of day or activity levels, and by facilitating real-time data collection techniques (e.g., EMA). For dynamic phenomena such as mood and depression, mobile technology could provide new standards of behavioural phenotypes against which clinical scales could be validated [28].

4.6.1. Limitations

A major limitation of this study is that the Whooley questions and the EPDS are both depression screening scales. Therefore, a score of 13 points or more on the EPDS does not indicate a diagnosis of depression. Therefore, to evaluate the true diagnostic properties of the Whooley questions, they need to be compared against a diagnostic gold standard as recently done by Littlewood and colleagues [27]. Nonetheless, the approach implemented in this study was chosen as it mirrors NICE recommendations for the recognition of antenatal depression [16], and in some clinics pregnant women are referred to mental health services on the basis of their EPDS scores.

4.6.2. Conclusion

This chapter showed that the scoring approach for the Whooley questions currently recommended by NICE (i.e., an affirmative answer to at least one Whooley question) may be appropriate for identifying pregnant women with EPDS scores above the scoring thresholds reported in the literature. In addition, this chapter suggests that a two-stage approach for the recognition of antenatal depression might be more appropriate than using a single case-finding instrument.

In the context of mHealth, this chapter demonstrated the supporting role that mobile technology can play for assessing if a given screening scale is suitable for specific clinical settings and for use with specific clinical populations. In addition, this chapter highlighted how consumer mobile devices could facilitate the assessment of clinical scales outside clinical settings, or the measurement of new behavioural phenotypes that could improve diagnostic accuracy. However, this rests on the assumption that patients are willing to install medical apps on their personal devices, and to adhere to a pre-specified survey administration protocol. This assumption needs to be tested. The study discussed in the next chapter evaluated the feasibility of using a bespoke mobile application for the monitoring of mood and depression throughout pregnancy in naturalistic settings through the use of the EPDS and EMA techniques. The key aim of this study was to measure patient adherence.

References (Chapter 4)

1. Stewart DE. Clinical practice. Depression during pregnancy. *New England Journal of Medicine*. 2011;365(17): 1605-11. Available from: doi: 10.1056/NEJMcp1102730 [Accessed 03 March 2017].
2. Mental Health Foundation. *Fundamental Facts About Mental Health 2015*. Available from: https://www.mentalhealth.org.uk/publications/fundamental-facts-about-mental-health-2015 [Accessed 03 March 2017].
3. Gavin NI, Gaynes BN, Lohr KN, Meltzer-Brody S, Gartlehner G, Swinson T. Perinatal depression: a systematic review of prevalence and incidence. *Obstetrics and Gynecology*. 2005; 106(5 Pt 1): 1071-83.
4. Lee AM, Lam SK, Sze Mun Lau SM, Chong CS, Chui HW, Fong DY. Prevalence, course, and risk factors for antenatal anxiety and depression. *Obstetrics and Gynecology.* 2007; 110(5): 1102-12.
5. Vigod SN, Brown S, Wilson CA, Howard LM. Depression in pregnancy. *BMJ*. 2016;352:i1547.
6. Bennett HA, Einarson A, Taddio A, Koren G, Einarson TG. Prevalence of depression during pregnancy: systematic review. *Obstetrics and Gynecology*. 2004;103(4): 698-709.
7. Austin MP, Middleton P, Reilly NM, Highet NJ. Detection and management of mood disorders in the maternity setting: the Australian Clinical Practice Guidelines. *Women and Birth*. 2013;26(1): 2-9. Available from: doi: 10.1016/j.wombi.2011.12.001 [Accessed 03 March 2017].
8. Heron J, O'Connor TG, Evans J, Golding J, Glover V, ALSPAC Study Team. The course of anxiety and depression through pregnancy and the postpartum in a community sample. *Journal of Affective Disorders*. 2004;80(1): 65-73.
9. Littlewood E, Ali S, Ansell P, Dyson L, Gascoyne S, Hewitt C, Keding A, Mann R, McMillan D, Morgan D, Swan K, Waterhouse B, Gilbody S, BaBY PaNDA study team. Identification of depression in women during pregnancy and the early postnatal period using the Whooley questions and the Edinburgh Postnatal Depression Scale: protocol for the Born and Bred in Yorkshire: PeriNatal Depression Diagnostic

Accuracy (BaBY PaNDA) study. *BMJ Open*. 2016;6(6): e011223. Available from: doi: 10.1136/bmjopen-2016-011223 [Accessed 03 March 2017].

10. Kinsella MT, Monk C. Impact of maternal stress, depression and anxiety on fetal neurobehavioral development. *Clinical Obstetrics and Gynecology*. 2009;52(3): 425-40. Available from: doi: 10.1097/GRF.0b013e3181b52df1 [Accessed 03 March 2017].
11. Pearson RM, Evans J, Kounali D, Lewis G, Heron J, Ramchandani PG, O'Connor TG, Stein A. Maternal depression during pregnancy and the postnatal period: risks and possible mechanisms for offspring depression at age 18 years. *JAMA Psychiatry*. 2013;70(12): 1312-9. Available from: doi: 10.1001/jamapsychiatry.2013.2163 [Accessed 03 March 2017].
12. Braithwaite EC, Murphy SE, Ramchandani PG. Effects of prenatal depressive symptoms on maternal and infant cortisol reactivity. *Archives of Women's Mental Health*. 2016; 19(4): 581-90. Available from: doi: 10.1007/s00737-016-0611-y [Accessed 03 March 2017].
13. Kingston D, Austin MP, Heaman M, McDonald S, Lasiuk G, Sword W, Giallo R, Hegadoren K, Vermeyden L, van Zanten SV, Kingston J, Jarema K, Biringer A. Barriers and facilitators of mental health screening in pregnancy. *Journal of Affective Disorders*. 2015;186: 350-7. Available from: doi: 10.1016/j.jad.2015.06.029 [Accessed 03 March 2017].
14. Howard LM, Molyneaux E, Dennis CL, Rochat T, Stein A, Milgrom J. Non-psychotic mental disorders in the perinatal period. *Lancet*. 2014;384(9956): 1775-88. Available from: doi: 10.1016/S0140-6736(14)61276-9 [Accessed 03 March 2017].
15. Gibson J, McKenzie-McHarg K, Shakespeare J, Price J, Gray R. A systematic review of studies validating the Edinburgh Postnatal Depression Scale in antepartum and postpartum women. *Acta Psychiatrica Scandinavica*. 2009;119(5): 350-64. Available from: doi: 10.1111/j.1600-0447.2009.01363.x [Accessed 03 March 2017].
16. National Collaborating Centre for Mental Health (2014 – 2015 Update), *Antenatal and postnatal mental health: clinical management and service guidance*. Clinical Guideline CG192.

17. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression – two questions are as good as many. *Journal of General Internal Medicine.* 1997;12:439-45.

18. Bosanquet K, Bailey D, Gilbody S, Harden M, Manea L, Nutbrown S, McMillan D. Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis. *BMJ Open*. 2015;5(12): e008913. Available from: doi: 10.1136/bmjopen-2015-008913 [Accessed 03 March 2017].

19. Mann R, Adamson J, Gilbody SM. Diagnostic accuracy of case-finding questions to identify perinatal depression. *CMAJ*. 2012;184(8): E424-30. Available from: doi: 10.1503/cmaj.111213 [Accessed 03 March 2017].

20. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Morrison C, Car J. Tablet computers for implementing NICE antenatal mental health guidelines: protocol of a feasibility study. *BMJ Open*. 2016;6(1): e009930. Available from: doi: 10.1136/bmjopen-2015-009930 [Accessed 03 March 2017].

21. Fawcett T. An introduction to ROC analysis. *Patter Recognition Letters*. 2006;27:861-874. Available from: doi: 10.1016/j.patrec.2005.10.010 [Accessed 29 July 2018].

22. Šimundić, AM. Measures of Diagnostic Accuracy: Basic Definitions. *The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine*. 2009;19(4):203-211.

23. Snap Surveys Limited. *Survey Software You'll Never Outgrow*. Available from: http://www.snapsurveys.com/survey-software/ [Accessed 01 March 2017].

24. Apple Inc. *iTunes Preview: Snap Mobile by Snap Surveys Limited.* Available from: https://itunes.apple.com/gb/app/snap-mobile/id507440809?mt=8 [Accessed 01 March 2017].

25. Snap Surveys Limited. *Snap WebHost – Online and Mobile Survey Management*. Available from: http://www.snapsurveys.com/survey-software/snap-webhost-online-mobile-survey-management-system/ [Accessed 01 March 2017].

26. R Core Team. *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing, Vienna, Austria; 2016. Available from: https://www.R-project.org/.

27. Littlewood E, Ali S, Dyson L, Keding A, Ansell P, Bailey D, Bates D, Baxter C, Beresford-Dent J, Clarke A, Gascoyne S, Gray C, Hackney L, Hewitt C, Hutchinson D, Jefferson L, Mann R, Marshall D, McMillan D, North A, Nutbrown S, Peckham E, Pervin J, Richardson Z, Swan K, Taylor H, Waterhouse B, Wills L, Woodhouse R, Gilbody S. Identifying perinatal depression with case-finding instruments: a mixed-methods study (BaBY PaNDA – Born and Bred in Yorkshire PeriNatal Depression Diagnostic Accuracy). *Health Services and Delivery Research*. 2018;6(6). Available from: doi: 10.3310/hsdr06060 [Accessed 28 December 2018].

28. Torous J, Baker JT. Why Psychiatry Needs Data Science and Data Science Needs Psychiatry: Connecting With Technology. *JAMA Psychiatry*. 2016;73(1): 3-4. Available from: doi: 10.1001/jamapsychiatry.2015.2622 [Accessed 03 March 2017].

5. Smartphones for the assessment and monitoring of antenatal depression in naturalistic settings using patients' own devices – Development of a study protocol

5.1. Overview

Consumer mobile devices, smartphones in particular, can enable the administration of self-reported mental health scales in naturalistic settings. These devices have also enabled a new model of data collection in which patients can install data collection apps directly onto their personal handsets. The characteristics of naturalistic settings combined with this new model of data collection offer the promise to increase the convenience of data collection for both patients and clinicians. In addition, this combination could enable the longitudinal collection of momentary and ecologically valid data that are better suited to the study of mental health phenomena such as antenatal depression.

This type of data collection however, is faced with new challenges that could affect data quality. One of these challenges is the level of patient adherence with systems that are intended for unsupervised use over a long period of time. Particularly since patient adherence is a prerequisite for data completeness and thus for the effectiveness of such systems. Therefore, their design should be aligned with the needs and motivation of their end users and should include features that enable them to achieve their goals.

This chapter and the next describe a study that evaluated the feasibility of using a bespoke app, BrightSelf, running on participants' own devices for the assessment of mood and depression during pregnancy. An important component in the development of this app and of the corresponding study protocol was the involvement of relevant stakeholders through a patient and public involvement (PPI) initiative. This initiative revealed the main motivating factors for pregnant women to use a system such as BrightSelf. It also highlighted the main challenges associated with the implementation of the proposed study protocol in antenatal services in the NHS.

This chapter discusses the key aspects of the study protocol that were informed by the findings of the PPI initiative. Namely, strategies for recruiting participants and for encouraging

them to install BrightSelf onto their handsets, the choice of self-reported scales, the types of assessments that were implemented, the design of an appropriate alerting system to ensure patient safety, and the choice of study outcomes. This chapter also discusses how this PPI initiative informed the way in which the primary outcome (i.e., participant adherence) was framed for the statistical analysis. The next chapter will present the results of the feasibility study and their interpretation.

5.2. Rationale

Apps running on consumer mobile devices can enable the ongoing assessment and monitoring of antenatal depression in naturalistic settings. In these settings, patient adherence is a key determinant of data completeness and of the effectiveness of an app. However, factors associated with this type of data collection could lead to poor patient adherence. These include the use of a personal device as a clinical tool, the demands on patients and/or clinicians, the potential reactivity associated with constant self-monitoring, and the compatibility of sampling protocols with existing care pathways. The potential impact of these factors on patient adherence needs to be assessed and accounted for when planning data collection strategies in order to minimise any detrimental effect on data quality (in this case, data completeness). This chapter and the next explore the feasibility of using a bespoke app running on patients' own devices for the assessment of antenatal depression throughout pregnancy in relation to patient adherence.

5.3. Background

As with other mental health conditions [1-4], consumer mobile devices could help to address some of the practical barriers to the screening and monitoring of antenatal depression. In particular, they can support the ongoing administration of screening scales and other self-reported measures in naturalistic settings. Data quality in these settings however, can be affected in ways that differ from clinical settings (as highlighted in Chapter 3). Therefore, this approach to data collection needs to be explored carefully before it is incorporated into antenatal care.

As discussed in Chapter 2, data collection in naturalistic settings is typically unsupervised and requires longer and more frequent sampling protocols. Similar to other electronic devices, consumer mobile devices have the technical capabilities to support this type of data collection. In addition, high levels of ownership have given rise to a promising approach that enables patients to use their own handsets for gathering clinical data [5, 6]. In this context, a potential advantage of this approach is that it can increase the convenience of antenatal depression screening and monitoring for both patients and clinicians. As shown in Chapter 3, approximately 98% of participants reported owning a smartphone. If this figure is

representative of the antenatal population, it suggests that practitioners could (after an initial resource-intensive app development phase) distribute a depression screening app to a large number of patients at a relatively low cost. Regular depression screening throughout pregnancy could then take place remotely, with further comprehensive clinical assessments and the provision of treatment being dependent on local referral and care pathways.

However, the use of patients' own devices for this purpose can present certain challenges that must be considered in order to prevent any potential detrimental effects on data quality. Namely, types of assessments, patients' willingness to use their personal handsets for clinical purposes, and the compatibility of this type of data collection with existing care pathways.

5.3.1. Retrospective versus momentary assessments: recall bias; sensitivity to temporal fluctuations; validity and reliability; and assessment reactivity

Depression screening in antenatal clinics usually relies on the use of validated scales (e.g., the Whooley questions and the EPDS) that ask patients to make retrospective assessments of their overall state across several symptom areas. Although these assessments can provide a valid and reliable indication of a woman's risk of antenatal depression, they are also susceptible to certain biases that could influence their accuracy and suitability for the study of mental health phenomena [7-11]. When completing retrospective assessments, patients tend to report previous states as being worse (or better) than they actually were, or to overestimate (or underestimate) the presence of certain symptoms [7, 11]. This phenomenon is known as recall bias. Research has also found that "recall" involves the reconstruction of past events with a person's current state acting as a reference point [7]. In addition, the ability of retrospective scales to capture fluctuations in a person's affect over time is very limited [11, 12].

The ubiquity and usage patterns of smartphones allow them to support the administration of self-report measures anywhere, at any frequency, and for any duration (at least from a technical point of view). Therefore, they are able to capture momentary, experiential, and ecologically valid data through techniques such as EMA. Being collected in real time, momentary data are less susceptible to the biases affecting retrospective scales [11-12]. In

addition, these data might be better suited to the study of mental health processes in real time, given their sensitivity to temporal fluctuations and to treatment response [11-13].

For these assessments to inform diagnostic or therapeutic decisions, they would first need to be validated against diagnostic procedures. Even if not used to inform clinical decisions directly due to lack of validation, momentary assessments could be administered alongside retrospective instruments such as the EPDS. One advantage of this approach is that it can help patients to make better sense of their condition. However, this approach can also lead to assessment reactivity. This phenomenon refers to the ability of momentary assessments to draw attention to the phenomenon of interest and thus influence its occurrence [13]. Assessment reactivity in depression is not well-understood, and it could have either a beneficial effect (by making patients more mindful of their actual experiences and decreasing the overestimation of symptoms) or a detrimental effect (by encouraging rumination of symptoms). For this reason, it needs to be explored.

Overall, choosing between retrospective and momentary assessments could be challenging for practitioners. Particularly when it comes to identifying a scale that is compatible with the characteristics of naturalistic settings and can also meet their needs as clinicians. For example, the overall assessments obtained through scales such as the EPDS offer the advantage of facilitating diagnostic or therapeutic decisions for non-specialist clinicians working in busy environments. This scale however, has not been validated for use in a multitude of non-clinical environments (e.g., at work, or when commuting) or at an increased frequency of administration.

5.3.2. Patients' willingness to use their personal handsets: adherence and data completeness

Another challenge associated with this approach to data collection is the willingness of patients to run a clinical app on their personal devices [6]. This factor could influence their adherence to sampling protocols (especially if they are clinician-led), and thus affect data completeness (i.e., the minimum amount of information required by healthcare practitioners to inform their decisions). This data quality domain is also susceptible to the demands that the intensity of sampling protocols (both in terms of their frequency and duration) can place

on patients, as well as on the value that patients might derive from gathering these data. Researchers and clinicians need to be mindful of these factors in order to plan data collection strategies that can take full advantage of the capabilities of consumer mobile devices.

5.3.3. Compatibility with existing services

This type of data collection can also affect the capacity and configuration of antenatal services. Ongoing antenatal depression screening through apps running on patients' own devices could result in a considerable increase in the number of data entries that healthcare providers would need to deal with. However, these providers might not have the capacity to review and action these entries promptly. Consequently, the implementation of data processing procedures that could notify clinicians when an action is required would need to be an essential feature of these apps. Alternatively, new dedicated data management roles or new communication channels between patients and clinicians might need to be introduced in clinical services as a way of dealing with the number of data entries.

This approach to antenatal depression screening might also result in antenatal services needing to become more proactive in their provision of mental health care (as long as they have the capacity to do so). For example, contacting patients that are at increased risk of antenatal depression before their next appointment is due. However, this needs to be balanced against the needs of patients and their privacy.

The impact of using patients' own devices for antenatal depression screening in naturalistic settings must be assessed in relation to clinical outcomes in the antenatal and postnatal periods, and in relation to its cost-effectiveness. However, the feasibility of this approach must be explored first to identify its potential impact on the quality of clinical data. This was the aim of this chapter: to evaluate the feasibility of using a bespoke mobile app running on patients' own devices for the assessment of mood and symptoms of depression throughout pregnancy. Primarily in relation to patient adherence.

A crucial step of this feasibility study was the engagement of relevant stakeholders in the research process through a PPI initiative. PPI sessions helped to inform key decisions about the study protocol and the bespoke mobile system [14]. In addition, the PPI sessions also revealed important differences between patients and healthcare providers in terms of their

motivation for using this bespoke system. These differences influenced how the primary outcome was analysed and interpreted. Therefore, the reporting of this study was divided into two chapters. The present chapter outlines the study protocol and the PPI initiative; it also discusses how the latter informed the study design. The next chapter presents the study findings and their interpretation.

5.4. Methods

The methods described in this section were published as a study protocol in an academic peer reviewed journal [15]. This study received favourable opinion from the South East Coast – Surrey National Research Ethics Committee on 21 December 2016, under the reference number 15/LO/0977. The study was implemented in NHS antenatal clinics across England between 03 April 2017 and February 2018.

The scope of this section is twofold: (i) to outline the study protocol as it was implemented in the participating NHS antenatal clinics; and (ii) to outline the PPI sessions that informed the study protocol and the design of the bespoke systems used in this study.

5.4.1. Protocol of the feasibility study

5.4.1.1. Aim

The aim of this study was to assess the feasibility of using a bespoke app, called BrightSelf, running on pregnant women's own devices for the assessment and monitoring of depression and mood during pregnancy. The primary outcome was patient adherence with a six-month sampling protocol. In addition, this study compared two types of assessments in relation to patient adherence: *Retrospective assessments*, and *Retrospective plus momentary assessments*.

5.4.1.2. Study design

Participants were randomly allocated to one of two types of assessment strategies:

- *Retrospective plus momentary assessments*: requiring the completion of (i) the EPDS; (ii) five momentary questions assessing the domains of mood, sleep, worry, enjoyment and energy levels; and (iii) two contextual questions once a month for six months
- *Retrospective assessments*: requiring the completion of the EPDS once a month for six months

5.4.1.3. Sample selection and participant recruitment

Participants were selected from women attending antenatal clinics in general practices, community services, and secondary care NHS centres in England during their first 14 weeks of pregnancy.

On the day of their antenatal appointment, potential participants were approached by a CSO or a research midwife, and were provided with a PIL. Potential participants expressing their interest in taking part in this study were assessed against the study's inclusion and exclusion criteria (Table 22).

Overall, recruitment took place in 13 localities across England: Blackburn, Burnley, Burton, Huntingdon, London, Manchester, Oldham, Rochdale, Stoke-on-Trent, Shrewsbury, Telford, Tyneside, and Wigan. A total of 350 women consented to take part in the study. However, two records had to be excluded as no participant ID number was recorded for them and it was not possible to identify their allocation and track app usage. Therefore, 348 women were randomised into the study: 172 women were allocated to the *Retrospective plus momentary assessments* group and 176 were allocated to the *Retrospective assessments* group.

Table 22. Participant inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
18 years old or older	Self-disclosed mental health problem at the time of being approached, or in the preceding 12 months
Up to 14 weeks pregnant, of any parity	Currently receiving treatment for any affective disorder
Attending antenatal care in the NHS	Not comfortable reading and writing in English
Own iPhone or Android smartphone	Incompatible handset (i.e., Windows phone, Blackberry or Linux powered smartphone) or not owning a smartphone

Potential participants meeting the inclusion criteria had all the study details explained to them and had the opportunity to ask as many questions about the study as they needed. Written, informed consent was obtained from those potential participants who, after receiving all the relevant information and having all questions answered to their satisfaction, still wished to take part in the study. Refusal to take part did not affect their legal rights, medical care or their relationship with their care providers.

After obtaining consent, participants were asked to self-complete a baseline assessment using a tablet computer. This assessment consisted of a socio-demographic survey, the Whooley questions, and the EPDS (Appendixes 32 to 34). Subsequently, the CSO or research midwife guided participants through the process of downloading and installing BrightSelf onto their own handsets. To activate the app, participants needed to enter a nine-digit activation code provided by the central research team. Participants were given detailed instructions on how to download and install BrightSelf at home (including entering the nine-digit activation code) if access to a wireless network was not available at the clinic and they did not wish to use their personal data allowances.

5.4.1.4. Surveys

5.4.1.4.1. Socio-demographic survey

A non-validated, 11-question survey was used to collect information about participants' age group, ethnic background, marital status, employment status, level of education, tablet computer ownership, obstetric history, and previous personal history of depression (Appendix 32).

5.4.1.4.2. Whooley questions

Participants were asked to complete the Whooley questions as per NICE guideline recommendations [16]. These two questions were developed as a case-finding instrument for depression in primary care, and assess depressed mood and anhedonia that have been present during the past month [17] (Appendix 33). Respondents were required to answer *Yes* or *No* to each question. In the presence of at least one affirmative answer to these questions, participants were allowed to continue participating in the study but were asked to see their midwife or GP to discuss any concerns.

5.4.1.4.3. Edinburgh Postnatal Depression Scale

The EPDS was also administered as part of this study. This is a ten-item self-administered survey that was originally developed to screen for postpartum depression [18] (Appendix 34). Since then it has been validated for use in the perinatal period and for use in community and clinical settings. This instrument assesses feelings of guilt, sleep disturbance, anhedonia and suicidal ideation that have been present during the past seven days. Each question is scored on a four-point scale ranging from zero to three points. An overall score is generated from the sum of these responses.

Although there is variability in the diagnostic accuracy of different EPDS scores, the following thresholds are commonly used: ten points for possible depression, 13 for probable depression, and 15 for antenatal depression [19]. In addition, special attention should be paid to item ten, as it deals with suicidal thoughts. Based on these scores, a clinician would be prompted to refer a woman to a mental health professional. The EPDS is a valid and reliable tool for identifying women who are at risk of depression, both during pregnancy and

postpartum. This instrument is also sensitive to changes in the severity of depression over time [18].

In the presence of an overall EPDS score of ten points or more, or one point or more on question ten, participants were allowed to take part in the study but were referred to their midwife or GP.

5.4.1.4.4. Momentary questions

Five non-validated momentary questions were used to assess participants' mood, sleep, worry, enjoyment and energy (Appendix 35). These questions were based on the work of a research fellow at the Collaboration for Leadership and Applied Health Research and Care (CLAHRC) for the East of England [20]. Each question was mapped onto five-point pictorial scales, ranging from one (low) to five points (high). Overall score calculations or the validation of these questions was outside the scope of this study.

5.4.1.4.5. Contextual questions

Two non-validated, contextual questions complemented the momentary mood questions (Appendix 36). They assessed participants' location and the activity in which they were engaged at the time they completed the five momentary questions.

5.4.1.4.6. Acceptance survey

Lastly, a non-validated, acceptance survey was administered at the end of the six-month participation period (Appendix 37). The purpose of these questions was to assess the acceptability to participants of BrightSelf, and to gather information about their experience of using it. These questions were derived from the Usefulness, Satisfaction and Ease of use (USE) questionnaire [21] and included a question about the desire to continue use (as a domain of user engagement [22]). This questionnaire also included questions about the experience of both the app and the self-reports.

5.4.1.5. BrightSelf

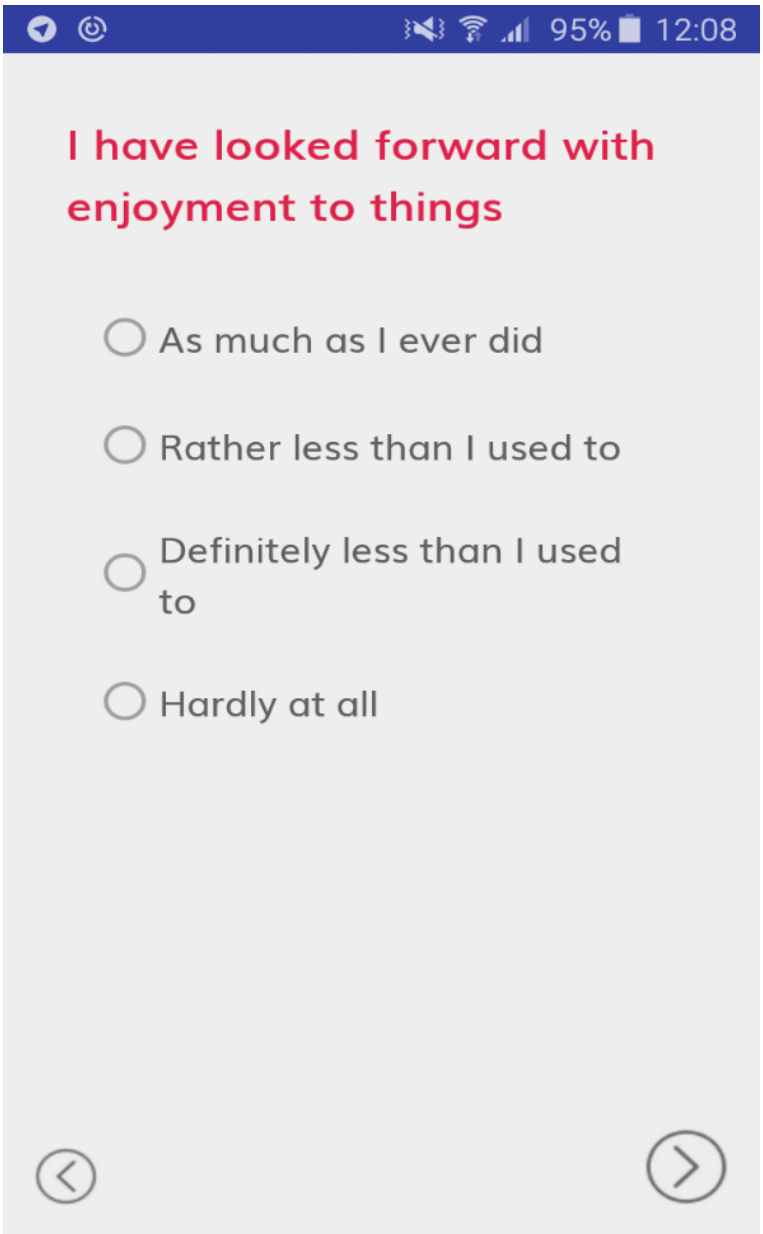
BrightSelf is a mobile system for the collection of self-reports during pregnancy. It elicited retrospective reports through the EPDS, and momentary reports according to the domains of

mood, sleep, worry, enjoyment and energy. These reporting features were labelled *Check Back* and *Check In*, respectively.

5.4.1.5.1. Check Back

This feature enabled the administration of the EPDS. It consisted of two introductory screens informing participants of the retrospective nature of this scale, and its intended clinical application. These screens were followed by the ten EPDS questions, presented one question at a time. Radio buttons were used to capture participants' responses (Figure 11).

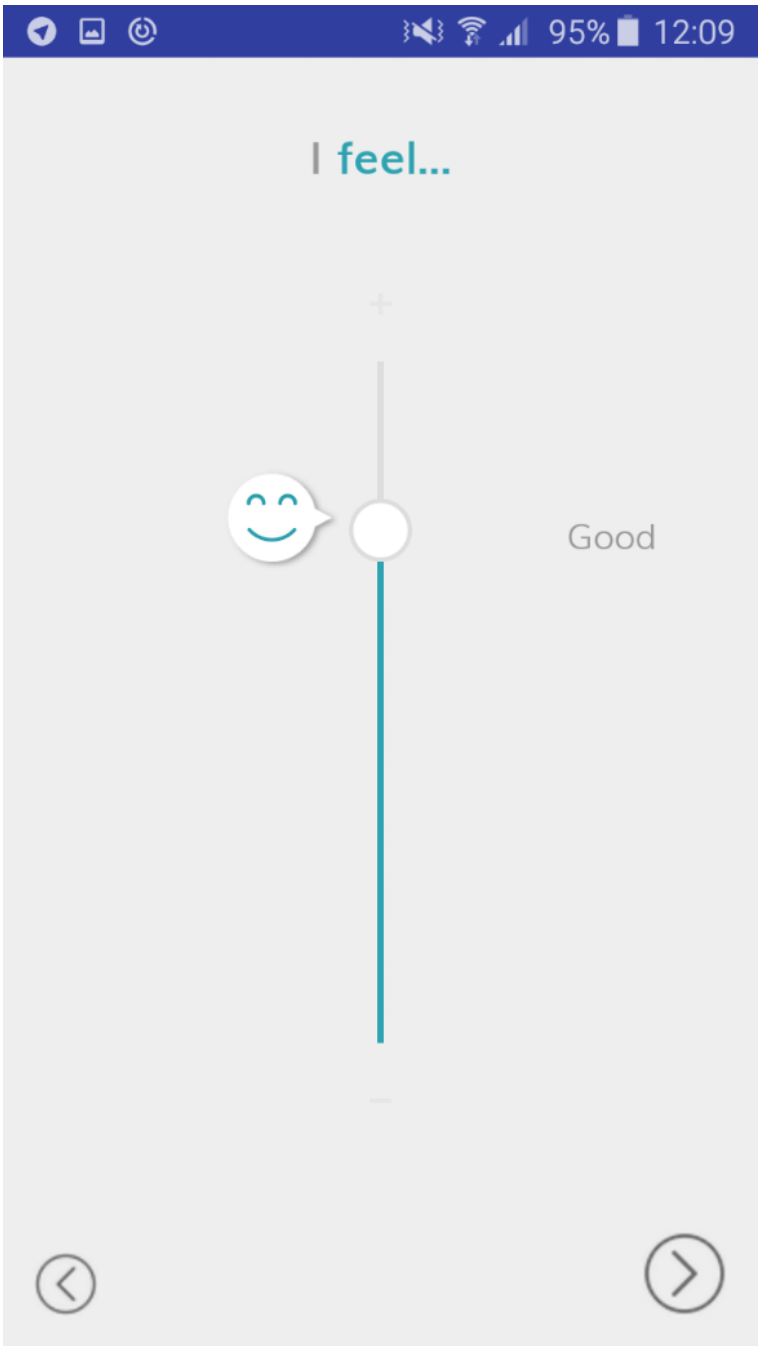
Figure 11. BrightSelf - Check Back (Screenshot)



5.4.1.5.2. Check In

This feature enabled the administration of the five momentary questions and the two contextual questions. The first five questions were presented using a five-point pictorial scale with temporary supporting text (see Figure 12). The last two questions used radio buttons to capture participants' responses.

Figure 12. BrightSelf - Check In (Screenshot)



5.4.1.6. Assessments

After completing the informed consent form, participants were asked to complete a baseline assessment at the clinic. This assessment consisted of the socio-demographic survey, the Whooley questions and an initial administration of the EPDS. Participants were then asked to complete monthly assessments at irregular intervals for six months.

Whenever an assessment was due (as per the sampling protocol), a notification was displayed on a participant's handset. There were two types of notifications: one for the *Check Back* feature and one for the *Check In* feature. Participants were able to respond to the notifications (and thus complete the corresponding assessments) or to dismiss them. If participants did not complete an assessment in response to a notification, they did not receive further reminders or follow-up notifications. In these instances, the non-response was recorded and coded as such.

Outside assessment periods, participants were able to use BrightSelf in any way they found convenient and there were no restrictions on the number of assessments they could complete on a single day.

At baseline, participants were randomly allocated to one of two types of assessment strategies: *Retrospective plus momentary assessments*, and *Retrospective assessments*.

5.4.1.6.1. Retrospective plus momentary assessments

As part of this assessment strategy, participants were asked to complete a combination of retrospective assessments and momentary assessments. A retrospective assessment was defined as a single administration of the EPDS. A momentary assessment was defined as the five momentary questions plus the two contextual questions.

Participants were required to complete assessments for a total of six months, and each assessment period consisted of six consecutive days. The six-month participation period was calculated from the day on which the app was activated (i.e., when participants entered their activation code and activated their account) and was divided into six intervals. All participants received the first notification two days after activating BrightSelf (i.e., first assessment), while

the use of the app was fresh within their minds and to give them an idea of what to expect during the study. Subsequent assessment periods were a random selection of six consecutive days within the 21 to 35 days following the end of the previous assessment period (in order to avoid 12 days of continuous assessments).

The assessment period was structured as follows:

- Day one: one retrospective assessment at any random time between 17:00 and 21:00
- Days two to five: three momentary assessments per day, displayed at random times within each of the following intervals: 09:00 – 12:00; 13:00 – 16:00; and 17:00 – 20:00
- Day six: one retrospective assessment at any random time between 17:00 and 21:00

At the end of the six-month period, participants were sent a link via SMS or by email to complete the non-validated, post-study acceptance survey.

5.4.1.6.2. Retrospective Assessments

As part of these assessments, participants were asked to complete retrospective assessments only. A retrospective assessment was defined as a single administration of the EPDS.

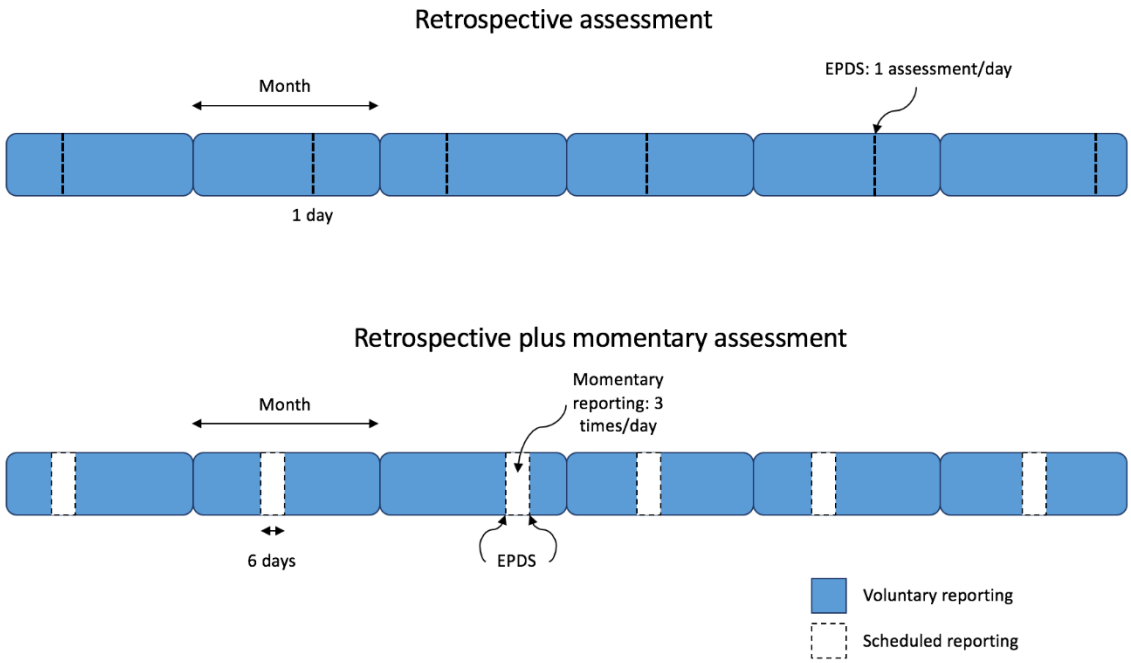
Participants were required to complete assessments for a total of six months, and each assessment period consisted of one day. The six-month participation period was calculated from the day on which the app was activated (i.e., when participants entered their activation code and activated their account) and was divided into six intervals. Participants received the first notification two days after activating BrightSelf (i.e., first assessment). Subsequent assessments took place on a random day within the 21 to 35 days following the end of the previous assessment period (in order to avoid two consecutive EPDS assessments).

On each assessment day, participants received a notification to complete the EPDS at a random time between 17:00 and 21:00.

At the end of the six-month period, participants were sent a link via SMS or by email to complete the non-validated, post-study acceptance survey.

See Figure 13 for an illustration of the assessment strategies implemented in this study.

Figure 13. Sampling protocols according to type of assessment strategy



5.4.1.7. Duty of care

Regardless of participants' allocation, the central research team was alerted if any of a series of pre-specified conditions were met (Table 23). These alerts were colour-coded according to their severity. In order to determine the level of severity, commonly-used EPDS scoring thresholds were chosen: 10 – 12 points for possible depression, and 13 points or more for probable depression. This approach has certain limitations, as the diagnostic accuracy of the EPDS can vary depending on the setting and the population to which it is administered.

Table 23. Alerting system

Type of alert	Criteria
Yellow (Mild)	Completing more than one EPDS report on the same day AND Overall EPDS score of 9 points or less AND Score of 0 on question 10 of the EPDS
Orange (Moderate)	Overall EPDS score between 10 and 12 points AND Score of 0 on question 10 of the EPDS
Red (Severe)	Overall EPDS score of 13 points or more OR Score of 1 point or more on question 10 of the EPDS

Before initiating the study, the central research team agreed a list of designated contacts with the relevant clinical care teams. These contacts included clinicians who were available during normal working hours, as well as those on duty outside working hours (including weekends and bank holidays). If a red or an orange alert was generated, the study coordinator contacted the designated member of the clinical care team by phone and email within 24 hours of receiving the alert. The clinical care team then followed up these alerts directly with the participants.

5.4.1.8. Randomisation

Block randomisation procedures with blocks of four were used in this study. Random numbers were generated using Stata 14 [23]. Each consecutive number was embedded within the nine-digit activation codes that were distributed to each participating NHS site. The full activation

codes were generated in the same order in which the random numbers were generated. Each NHS site received a list of activation codes on a first-come-first-served basis. Members of the local research teams were not informed of the random sequence generation and were not able to identify which digit determined participant allocation. They were required to use activation codes sequentially, in the same order in which participants were recruited. Activation codes referring to the *Retrospective plus momentary assessments* condition activated a version of BrightSelf in which both the *Check Back* and the *Check In* features were active. Activation codes referring to the *Retrospective assessments* condition activated a version of BrightSelf in which only the *Check Back* feature was active.

5.4.1.9. Outcomes

5.4.1.9.1. Adherence to sampling protocols

Multiple definitions of this outcome were used:

a) Number of participants who completed 100% of the expected assessments (as per the study protocol) as a proportion of the total number of participants who were randomised into the study
b) Number of participants who completed at least one expected assessment in each sampling period for all six sampling periods
c) Number of participants who completed at least one assessment during the entire duration of the study

5.4.1.9.2. Drop-out rates

Number of participants who did not complete the six-month participation period as a proportion of the total number of participants who were randomised into the study.

5.4.1.9.3. Acceptance

Participants' ratings and responses to the post-study acceptance survey.

5.4.1.9.4. Alerts

Number and type of alerts that were generated throughout the study. The number of participants who triggered alerts was also calculated.

The details of the statistical analyses are presented in Chapter 6.

5.4.2. Patient and public involvement strategy

The development of this protocol involved a PPI initiative, which consisted of a series of sessions with relevant stakeholders: users of antenatal services, clinicians and academics working in the field. The methods of (and finding from) this initiative were published in a peer-reviewed academic journal [14].

5.4.2.1. Aim

The aim of this PPI initiative was to understand the psychological wellbeing needs of pregnant women, to understand the motivation of parents and healthcare professionals for using mobile technology during the perinatal period, and to gather feedback on an early prototype of BrightSelf.

5.4.2.2. Structure of the PPI sessions

The PPI sessions involved five focus groups and 17 one-to-one semi-structured interviews, all of which followed the same format. At the beginning of each session, the researchers introduced themselves and their research work. Participants were then encouraged to introduce themselves, to share their experience of using mobile apps during pregnancy, and to share their interest in perinatal mental health. The latter was entirely optional as participants' interest in this area might have been triggered by their own personal experience of mental health problems. Thereafter, the researchers initiated a discussion around the psychological wellbeing needs of parents during the perinatal period, and the role of antenatal services and mobile technology in supporting these needs. Participants were then asked to generate ideas on the potential use of mobile technology for the assessment and monitoring of depression during pregnancy. They were subsequently shown a prototype of BrightSelf and were asked for their feedback. Lastly, the researchers disclosed the intended use of BrightSelf as part of this feasibility study, and asked participants to provide feedback on the study protocol.

5.4.2.3. Participants

Overall, 38 participants took part in the PPI sessions: 15 midwives, one CSO, one psychologist, six clinical researchers in the area of maternal mental health, two general practitioners, one clinical psychologist, one psychiatrist, eight pregnant women, and three women who had recently given birth. None of the participants had ever been diagnosed with perinatal depression or any other common mental health disorder.

5.5. Results

The results of this feasibility study are presented in Chapter 6. With regard to the PPI sessions, they were part of a collaboration with a PhD candidate at Trinity College Dublin, who focused on human computer interaction factors in the design of mobile systems for the collection of self-reported data. Therefore, the findings from these sessions are not presented here as their primary purpose was to inform the design of BrightSelf from a human computer interaction perspective. As such, these findings are outside the scope of this PhD thesis.

5.6. Discussion

Although the primary purpose of the PPI sessions was to inform the design of BrightSelf, they helped to inform five aspects of the study protocol: participant recruitment; choice of self-reported scales; types of assessments; the duty of care protocol; and the choice of study outcomes. In addition, the PPI sessions also revealed important differences between pregnant women and healthcare professionals in terms of their motivation for using a system such as BrightSelf. These differences raised important questions about the concept of patient adherence, which informed how this outcome was analysed and interpreted. Therefore, discussing these themes is necessary to contextualise the study findings presented in Chapter 6.

5.6.1. Participant recruitment

The main decisions about participant recruitment were the limit on gestational age and the strategy for asking participants to download BrightSelf onto their handsets.

Participants were recruited during their booking appointment to ensure that they were in their first pregnancy trimester, or as close as possible to it. As mentioned in Chapter 1, this is the first official antenatal appointment and is supposed to take place at around week ten. In practice, it usually occurs between 12 and 14 weeks. For this reason, the limit on gestational age was set at 14 weeks.

There were two key determinants for recruiting participants at this stage. Firstly, the PPI sessions revealed that the booking appointment is the longest appointment, and the one where patients and their midwifes have the opportunity to discuss issues other than the physical health of the mother in relation to foetal development. In this context, discussing an app for the assessment of mood and depression throughout pregnancy would not feel out of context. Moreover, the booking appointment would give midwives enough time to guide participants through the process of downloading and installing BrightSelf onto their handsets, as subsequent antenatal appointments are considerably shorter (approximately 15 minutes).

The second reason was to ensure that most assessments (if not all) took place whilst women were receiving care from their antenatal care providers, as this feasibility study was concerned with antenatal depression. The intention was also to minimise the risk of not being able to implement the duty of care protocol due to changes in healthcare providers after childbirth.

Concerning the installation of BrightSelf, the PPI sessions revealed that the best approach would be to ask participants to download the app in the clinic under the guidance of a midwife or CSO. Those who took part in the PPI sessions believed that this approach would (i) help to set the context in which BrightSelf was used, (ii) ensure participant adherence with this initial ask, and (iii) create an environment where participants could familiarise themselves with the main features of the app and clarify any doubts. In addition, and in accordance to existing literature [24], some PPI stakeholders suggested that endorsement by clinicians could promote patients' long-term engagement with an app.

As a result of this, midwives were expected to become proficient users of BrightSelf before being able to guide participants. This was supported by the multiple requests received from local research teams to download and test BrightSelf before the start of the study. For

research midwives, this fell within their research duties. For purely clinical midwives however, this had implications for their workload as they had to familiarise themselves with an app that they would not normally use in clinical practice. One challenge of this approach that was highlighted during the PPI sessions (and later confirmed by local research teams) was the need for clinical services to provide the basic infrastructure for the distribution of BrightSelf. In this case, providing access to Wi-Fi in the clinics, particularly for patients with limited or pay-as-you-go data plans.

5.6.2. Choice of self-reported scales

The decision to incorporate momentary questions was to assess if they could influence adherence with the sampling protocol. Practitioners taking part in the PPI sessions expressed the need for an instrument that could facilitate decisions about patient care. For this reason, the EPDS was included in BrightSelf for all participants, as per NICE recommendations [16]. Pregnant women and women who had recently given birth also acknowledged how the inclusion of this scale might be useful for their clinicians and might be relevant in the context of antenatal care. However, they also highlighted that their motivation for using a system such as BrightSelf would be to facilitate self-reflection, and to enable comparisons with normative experiences of mood changes during pregnancy. Momentary questions were perceived as being better able to satisfy these needs, particularly since the domains that were considered resonated with their experiences. This prompted the use of momentary questions as a strategy to promote effective engagement with BrightSelf.

5.6.3. Momentary and retrospective assessments

The main decisions concerning the implementation of the momentary and retrospective assessments were the initiation, the frequency and duration of the sampling protocols, and the repeated administration of the EPDS to participants in the *Retrospective plus momentary assessments* group.

Regardless of their allocation, all participants received the initial notification two days after activating BrightSelf. The reason behind this decision was to encourage interaction with the app as soon as possible, while the demonstration given by midwives was still fresh in participants' minds. In addition, this was intended to give participants a sense of the

commitment required for this study. It was thought that this decision could influence long-term adherence with the sampling protocol, and thus data completeness.

Concerning the frequency and duration of the sampling protocols, there was no available baseline data to inform this decision. The monthly frequency was predominantly dictated by the intention to avoid successive administrations of the EPDS that overlapped in the time frame they covered. The six-month duration of the sampling protocol was informed by the expected duration of the interaction between pregnant women and antenatal services given that they were recruited during their booking appointment.

The decision to administer the EPDS at each tail-end of the momentary assessments was to determine the presence of assessment reactivity.

5.6.4. Alerting system and duty of care protocols

The aim of the alerting system and the duty of care protocol was to explore the data management challenges associated with a system such as BrightSelf. During the PPI sessions, many practitioners highlighted the potentially high volume of data points that would be generated by each participant, and the inability of clinicians to review all of them. Therefore, it was necessary to implement a system that would generate alerts whenever a set of conditions based on clinical criteria were met. Since the data management demands were unknown, the central research team agreed to be responsible for this aspect of the alerting system. Particularly, assessing the accuracy of the alerts. However, clinical teams were responsible for the clinical aspect of the alerts given their duty of care towards their patients.

5.6.5. Study outcomes

The primary outcome of this feasibility study was patient adherence to the sampling protocol. This choice reflects one of the issues at the core of systems that are intended for long-term, unsupervised use: patient engagement with these systems is a pre-requisite to their effectiveness [25]. Existing consumer mobile devices can handle the technical requirements of data gathering, data processing and data communication. However, the systems or apps that are developed to run on these devices ought to be sensitive to the needs, motivations and routines of patients, and help them to achieve their goals [25]. This has been proposed

as one of the key strategies to improve long-term engagement with these systems [25, 26]. Otherwise, poor rates of engagement could threaten data completeness and affect the clinical decisions that rely on these data.

5.6.6. Notions of patient adherence

The differences between pregnant women and healthcare practitioners in their motivation for using a system such as BrightSelf raised important questions about perceived notions of patient adherence. Current conceptual frameworks recognise the active role of patients in their own healthcare [27]. As such, these frameworks assume that in order to reach a therapeutic decision, patients and their healthcare providers engage in a meaningful dialogue or in a negotiation process as equal partners [27]. In this context, the role of healthcare providers is predominantly to offer guidance and advice on the basis of their expert knowledge.

Conceptual frameworks relevant to self-monitoring and self-tracking, propose that digital health (including mHealth) could help patients to undergo a process of self-discovery, either on their own or in collaboration with their healthcare providers [28]. These journeys of self-discovery are perceived as individual experiences that can lead to multiple patterns of adherence [28].

These frameworks promote a personalised view of adherence that, in this study, could have manifested itself as a discussion between pregnant women and their midwives about the appropriateness of monitoring mood and depression, the relevance of the momentary domains, and the optimal frequency and duration of this process. However, resourcing, methodological and governance constrains prevented the research team from implementing this approach.

Instead, participants were required to complete monthly assessments of the EPDS and of mood, sleep, worry, enjoyment and energy for six months. This approach, which is still prevalent in the healthcare literature, is more compatible with traditional notions of adherence according to which patients are expected to unquestioningly comply with the requests of researchers or healthcare practitioners [27]. Therefore, the measure that was

proposed in the original study protocol [15] (i.e., number of participants who completed all the expected assessments) provided an indication of patient compliance rather than of patient adherence. For this reason, two definitions that were deemed more compatible with the latter were added to the analysis.

Overall, the following definitions were considered:

a) Number of participants who completed the total number of expected assessments
b) Number of participants who completed at least one expected assessment in each sampling period for all six sampling periods
c) Number of participants who completed at least one assessment for the entire duration of the study

By considering these definitions, it was possible to discuss important issues about perceived notions of adherence in relation to mHealth, which are presented in Chapter 6.

5.6.7. Conclusion

The PPI sessions revealed important differences between pregnant women and antenatal care providers in their motivation for using a system such as BrightSelf. These differences informed key aspects of the study design: approach to participant recruitment; choice of self-reported scales; type of assessments; the design of the alerting system; and the choice of study outcomes. In relation to the latter, patient adherence was identified as the primary outcome of this feasibility study as it is a pre-requisite for the effectiveness of mHealth interventions. However, the tension between pregnant women and their healthcare providers in terms of their goals and needs highlighted the need to explore different patterns of adherence. These are presented and discussed in the next chapter.

References (Chapter 5)

1. BinDhim NF, Shaman AM, Trevena L, Basyouni MH, Pont LG, Alhawassi TM. Depression screening via a smartphone app: cross-country user characteristics and feasibility. *Journal of the American Medical Informatics Association (JAMIA).* 2015;22(1): 29-34. Available from: doi: 10.1136/amiajnl-2014-002840 [Accessed 12 May 2017].
2. BinDhim NF, Alanazi EM, Aljadhey H, Basyouni MH, Kowalski SR, Pont LG, Shaman AM, Trevena L, Alhawassi TM. Does a Mobile Phone Depression-Screening App Motivate Mobile Phone Users With High Depressive Symptoms to Seek a Health Care Professional's Help? *Journal of Medical Internet Research*. 2016;18(6): e156. Available from: doi: 10.2196/jmir.5726 [Accessed 12 May 2017].
3. Woo YS, Bahk WM, Hong J, Yoon BH, Hwang TY, Kim MD, Jon DI. Use of a smartphone application to screen for bipolar sprectrum disorder in a community sample. *Health Informatics Journal*. 2016;22(3): 779-88. Available from: doi: 10.1177/1460458215589601 [Accessed 12 May 2017].
4. Torous J, Chan SR, Tan SYM, Behrens J, Mathew I, Conrad EJ, Hinton L, Yellowlees P, Keshavan M. Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics. *JMIR Mental Health*. 2014;1(1): e5. Available from: doi: 10.2196/mental.4004 [Accessed 12 May 2017].
5. Coons SJ, Eremenco S, Lundy JJ, O'Donoghue P, O'Gorman H, Malizia W. Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials. *Patient.* 2015;8(4): 301-309. Available from: doi: 10.1007/s40271-014-0090-z [Accessed on 12 May 2017].
6. Pugliese L, Woodriff M, Crowley O, Lam V, Sohn J, Bradley S. Feasibility of the "Bring Your Own Device" Model in Clinical Research: Results from a Randomized Controlled Pilot Study of a Mobile Patient Engagement Tool. *Cureus*. 2016;8(3):e535. Available from: doi: 10.7759/cureus.535 [Accessed 12 May 2017].
7. Blome C, Augustin M. Measuring change in quality of life: bias in prospective and retrospective evaluation. *Value in Health*. 2015;18(1): 110-5. Available from: doi: 10.1016/j.jval.2014.10.007 [Accessed 12 May 2017].

8. Moskowitz DS, Young SN. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience*. 2006;31(1): 13-20.
9. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. *Annual Review of Clinical Psychology*. 2008;4:1-32.
10. Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study. *Journal of Medical Internet Research*. 2016;18(3): e72. Available from: doi: 10.2196/jmir.5505 [Accessed on 12 May 2017].
11. Ben-Zeev D, McHugo GJ, Xie H, Dobbins K, Young MA. Comparing retrospective reports to real-time/real-place mobile assessments in individuals with schizophrenia and a nonclinical comparison group. *Schizophrenia Bulletin*. 2012;38(3): 396-404. Available from: doi: 10.1093/schbul/sbr171 [Accessed 12 May 2017].
12. Solhan MB, Trull TJ, Jahng S, Wood PK. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. *Psychological Assessment*. 2009;21(3): 425-36. Available from: doi: 10.1037/a0016869 [Accessed 12 May 2017].
13. Van Ballegooijen W, Ruwaard J, Karyotaki E, Ebert DD, Smit JH, Riper H. Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): protocol of a randomised controlled trial. *BMC Psychiatry*. 2016;16: 359. Available from: doi: 10.1186/s12888-016-1065-5 [Accessed 14 September 2018].
14. Doherty K, Barry M, Marcano-Belisario JS, Arnaud B, Morrison C, Car J, Doherty G. A Mobile App for the Self-Report of Psychological Well-Being During Pregnancy (BrightSelf): Qualitative Design Study. *JMIR Mental Health*. 2018;5(4): e10007. Available from: doi: 10.2196/10007 [Accessed 30 December 2018].
15. Marcano-Belisario JS, Doherty K, O'Donoghue J, Ramchandani P, Majeed A, Doherty G, Morrison C, Car J. A bespoke mobile application for the longitudinal assessment of depression and mood during pregnancy: protocol of a feasibility study. *BMJ Open.* 2017;7: e014469. Available from: doi: 10.1136/bmjopen-2016-014469 [Accessed 24 June 2018].

16. National Collaborating Centre for Mental Health (2014 – 2015 Update), *Antenatal and postnatal mental health: clinical management and service guidance*. Clinical Guideline CG192.
17. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression – two questions are as good as many. *Journal of General Internal Medicine.* 1997;12:439-45.
18. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *British Journal of Psychiatry.* 1987; 150:782-786.
19. Gibson J, McKenzie-McHarg K, Shakespeare J, Price J, Gray R. A systematic review of studies validating the Edinburgh Postnatal Depression Scale in antepartum and postpartum women. *Acta Psychiatrica Scandinavica*. 2009;119(5): 350-64. Available from: doi: 10.1111/j.1600-0447.2009.01363.x [Accessed 03 March 2017].
20. Blake F. *Mumshine-The App: The design of a micro on-line intervention for postnatal women to improve mood and well-being using an app for the smart phone*. [Lecure] CLAHRC East of England Fellows Showcase Event. 04 December 2014.
21. Lund AM. Measuring usability with the USE questionnaire. *Usability Interface*. 2001;8(2): 3-6.
22. Schoenau-Fog H. Hooked!-evaluating engagement as continuation desire in interactive narratives. *Proceedings of the 4th International Conference on Interactive Digital Storytelling*, 28 November-01 December 2011, Berlin, Heidelberg. Available from: doi: 10.1007/978-3-642-25289-1_24 [Accessed 12 May 2017].
23. StataCorp. *Stata Statistical Software: Release 13* [Software]. College Station, Texas: StataCorp LP; 2013.
24. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC Medical Informatics and Decision Making*. 2016;16(1): 120. Available from: doi: 10.1186/s12911-016-0359-3 [Accessed 12 May 2017].

25. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, Merchant GC, Naughton F, Blandford A. Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *Americal Journal of Preventive Medicine.* 2016;51(5): 833-842. Available from: doi: 10.1016/j.amepre.2016.06.015 [Accessed 23 July 2018].

26. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence-Based Mental Health*. 2018; 21(3): 116-119. Available from: doi: 10.1136/eb-2018-102891 [Accessed 17 September 2018].

27. Ahmed R, Aslani P. What is patient adherence? A terminology overview. *International Journal of Clinical Pharmacology*. 2014;36:4-7.

28. Neff G, Nafus D. *Self-Tracking*. London, UK: MIT Press; 2016.

6. Smartphones for the assessment and monitoring of antenatal depression in naturalistic settings using patients' own devices – Study findings

6.1. Overview

This chapter is a continuation of Chapter 5, presenting the findings of a study that explored the feasibility of using a bespoke app running on pregnant women's own smartphones for the assessment and monitoring of depression during pregnancy through the EPDS. The primary outcome of this study was adherence with a pre-specified protocol, as this outcome has been identified as a key determinant of the effectiveness of mHealth systems. In addition, this study evaluated the effect of momentary questions on the levels of adherence that were observed. This approach was informed by the findings of a PPI strategy, which revealed that the main motivation for pregnant women to use a system such as the one evaluated in this chapter was to self-reflect and to make comparisons with normative mood changes during pregnancy. These momentary questions were perceived as being ideal for supporting these goals. Moreover, three definitions of adherence were used to explore some of the challenges associated with the assumptions that underpin the development of mHealth interventions. Other outcomes were considered as indicators of the implementation challenges associated with the deployment of this type of technology in antenatal clinics.

Similar to what has been reported in the literature, the findings of this study revealed poor levels of participant adherence with the bespoke app as most commonly measured. Potential explanations include the heterogeneity of pregnant women in terms of their psychological wellbeing needs, the inability of the app to satisfy the needs of its intended audience, or the de-contextualisation of the app when deployed as part of a research study. Nonetheless, this study provided an excellent opportunity to explore the different patterns of participant adherence that were observed. This highlights the need to understand adherence as a cluster of individualised patterns rather than as a summative process. In doing so, there is a better chance to devise strategies that can successfully encourage effective engagement with mHealth interventions.

6.2. Rationale

As outlined in Chapter 5, apps running on consumer mobile devices can enable the ongoing assessment and monitoring of antenatal depression in naturalistic settings. In these settings, patient adherence has been identified as one of the key determinants of data completeness and of the effectiveness of an app. This conceptualisation of adherence however, is still heavily informed by traditional notions of healthcare. This can lead to the development of mHealth systems that are at odds with the way patients experience health-related phenomena, thus creating unrealistic expectations about their willingness to engage with these systems. This also continues to promote a one-size-fits-all model of healthcare provision, which is opposed to the idea of mHealth as an enabler of personalised care. Instead, adherence may be better conceptualised as a complex phenomenon that varies both between and within individuals. Mobile apps may prove useful in understanding this complexity, which is essential for the design of mHealth systems that are better aligned with patients' experiences and can better support their care. This chapter presents the results of a study that explored these themes by evaluating the feasibility of using a bespoke app running on patients' own devices for the assessment of antenatal depression throughout pregnancy.

6.3. Background

This study assessed the feasibility of using a bespoke mobile system (i.e., BrightSelf) running on participants' own smartphones for monitoring mood and symptoms of depression during pregnancy. As outlined in Chapter 5, the main study outcome was participant adherence with the sampling protocol. Sometimes referred to as engagement, adherence has been identified as a pre-requisite for the effectiveness of mHealth systems that are intended for unsupervised use over a prolonged period of time [1]. In the context of the data quality dimensions outlined in Chapter 2 [2], adherence with a sampling protocol can affect the completeness of clinical data, and hence the diagnostic and/or therapeutic decisions informed by these data. For example, when identifying pregnant women who are at increased risk of antenatal depression.

Two aspects were central to the assessment of adherence in this feasibility study: EMA as a strategy to promote adherence; and the use of multiple definitions of participant adherence.

6.3.1. Ecological momentary assessments as a strategy to promote participant adherence

Participant-led co-design (or co-creation) has been identified as a useful strategy to promote adherence with mHealth systems [1, 3]. This process can help to identify the needs and goals of users, and to ensure that they are incorporated in the design of these systems. That was the purpose of the PPI initiative that complemented this feasibility study, which revealed that the main motivation for pregnant women to use a system such as BrightSelf was to facilitate self-reflection, and to enable comparisons with normative experiences of mood changes during pregnancy [4]. Compared to the EPDS, momentary assessments were perceived as being better able to meet these needs. Consequently, one group of participants was randomly allocated to a version of BrightSelf that elicited momentary assessments in addition to the EPDS assessments that all participants were required to complete. The objective was to determine if these momentary assessments had any impact on the levels of adherence with BrightSelf. All participants were required to complete the EPDS as this scale directly informs decisions about the provision of mental health care in antenatal services. Thus, it satisfies the needs of healthcare practitioners as the other user group of BrightSelf.

6.3.2. Multiple definitions of participant adherence

The PPI initiative also raised questions about notions of adherence. Current conceptualisations of this construct assume that patients play an active role in making decisions about their health or the care they receive [5]. However, research governance requirements led to the research team having to prescribe the frequency and duration of the sampling protocol. This situation resembles traditional notions of adherence in which patients are expected to comply with requests imposed on them on the basis of professional experience [5]. Therefore, three definitions of adherence were considered as they enabled the assessment of aspects that were compatible with the different notions of this outcome.

The first definition of adherence considered the total number of participants who completed all the expected assessments for the entire duration of the study. This definition is consistent

with traditional models of adherence, in which participants were expected to complete 100% of all assessments as determined by the random notifications. The second definition of adherence considered the number of participants who completed at least one expected assessment in each sampling period for the entire duration of the study. This definition adopted more flexible criteria for considering participants as adherent, as it acknowledged that women in the *Retrospective plus momentary assessments* group had more opportunities to complete an assessment than participants in the *Retrospective assessments* group. The challenge in this situation was the tension between the needs of clinicians and the needs of patients as two distinct user groups of BrightSelf. For example, the completion of momentary assessments did not assist clinicians in their decision-making.

Lastly, the third definition of adherence considered the number of participants who completed at least one expected assessment at any point during the study. This definition is more compatible with current notions of adherence, as it acknowledges that participants might not have followed the pre-specified sampling protocol in its entirety. Yet they might have used BrightSelf in any way that suited their needs. Although the tension between the needs of clinicians and the needs of patients was also relevant in this instance, the main issue became about the need to consider adherence as a cluster of different patterns of individualised use rather than as a summative process [6]. This then prompted an exploratory analysis of assessment completion that was independent from the sampling protocol.

Overall, this chapter evaluated the effect of two assessment strategies (i.e., *Retrospective plus momentary assessments*, and *Retrospective assessments*) on participant adherence with BrightSelf, by considering multiple definitions of this outcome. In addition, this chapter also evaluated the effect of these strategies on other outcomes that are related to the implementation of mHealth systems in naturalistic settings: drop-out rates; acceptance to participants; and the number of high-risk alerts that were triggered.

6.4. Methods

Full details of the study design and the methods of this feasibility study were presented in Chapter 5 and published in a peer reviewed journal [7]. This section presents a summary of the study design, the study outcomes and the statistical analyses.

6.4.1. Study design

Three hundred and fifty pregnant women of up to 14 weeks of gestational age were recruited from antenatal clinics in 13 localities across England. Participants were asked to complete a baseline assessment consisting of the Whooley questions, the EPDS and a socio-demographic survey. Participants were then asked to download and install BrightSelf on their own smartphones. This app elicited retrospective assessments through the EPDS, and momentary assessments through the domains of mood, sleep, worry, enjoyment and energy.

Participants were randomised to one of two assessment strategies: *Retrospective plus momentary assessments* or *Retrospective assessments*. The former consisted of monthly six-day assessment periods during which participants had to complete the EPDS and momentary questions. *Retrospective assessments* consisted of monthly one-day assessment periods during which participants were asked to complete the EPDS. Participation in the study lasted for six months, at the end of which participants were asked to complete a post-study acceptance survey. Figure 13 (in Chapter 5) shows a visual representation of the assessment strategies that were implemented in this feasibility study.

An alerting system notified the central research team whenever a participant met clinical criteria for high risk of antenatal depression. The research team then notified the clinical team responsible for the care of that patient so that they could take appropriate action.

6.4.2. Sample size

The proposed sample size was related to the 95% confidence interval for the adherence rate, as recommended by the Research Design Service in London. Therefore, 96 participants were needed in each arm with a 95% confidence level and a confidence interval of ten. This translated into a total sample of 192 (i.e., 96 participants in each experimental group). For this reason, the aim was to recruit at least 250 participants, to account for drop-outs and potential miscarriages.

6.4.3. Study outcomes

6.4.3.1. Adherence to sampling protocols

As outlined in the background section and in Chapter 5, multiple definitions of this outcome were used:

a) Number of participants who completed 100% of the expected assessments (as per the study protocol) as a proportion of the total number of participants who were randomised into the study
b) Number of participants who completed at least one expected assessment in each sampling period for all 6 sampling periods
c) Number of participants who completed at least one assessment during the entire duration of the study

An assessment was considered to have been completed in response to a notification if it occurred before the next notification was due.

Moreover, the third definition of adherence prompted an exploratory analysis of individual patterns of assessment completion. As this analysis was independent from the sampling protocol, the duration of the participation period was divided into six one-month intervals starting from the date of the first BrightSelf notification. This analysis considered the number of intervals for which participants completed assessments, as well as the combination of these intervals.

6.4.3.2. Drop-out rates

Number of participants who did not complete the six-month participation period as a proportion of the total number of participants who were randomised into the study.

6.4.3.3. Acceptance

Participants' ratings and responses to the post-study acceptance survey.

6.4.3.4. Alerts

Number and type of alerts that were generated throughout the study. The number of participants who triggered alerts was also calculated.

6.4.4. Statistical analyses

6.4.4.1. Descriptive statistics

For each experimental group, the following information was reported:

- Demographic characteristics (as captured by the non-validated, socio-demographic survey)
- Proportion of participants answering *Yes* to any of the Whooley questions during the baseline assessment
- Proportion of participants scoring at each interval of the EPDS during the baseline assessment: between 0 and 9 points; between 10 and 12 points; and 13 points or more
- Proportion of participants scoring one or more points on question ten of the EPDS during the baseline assessment

6.4.4.2. Inferential statistics

The *Retrospective and Momentary assessments* and the *Retrospective assessments* groups were compared for differences in adherence rates, drop-out rates, number of alerts triggered, and number of participants who triggered alerts. Chi-squared tests of independence, or Fisher exact tests of independence were used for these comparisons.

Usage of BrightSelf was presented descriptively as the different patterns that emerged in relation to the intervals for which participants completed assessments. Similarly, post-study acceptance was presented descriptively by assessing the proportion of participants who selected each response option on a 7-point Likert scale for each of the questions in the post-study acceptance survey. This approach was adopted because there was no way of ensuring that respondents interpreted the concepts explored by this survey in the same way.

The intention was to evaluate participants' momentary and retrospective assessments through linear mixed effects models [8, 9], which consider between group as well as within

participant differences over time. However, this was not possible due to the high levels of missing data and the violation of the missing at random assumption. Similarly, the intention was to evaluate a potential effect of the momentary assessments on the retrospective assessments through a linear mixed effects model but this was not possible for the reasons outlined above.

6.5. Results

6.5.1. Baseline assessments

6.5.1.1. Characteristics of the participants

Table 24 presents a summary of participants' socio-demographic characteristics. The majority of participants were between 28 and 32 years of age, white, married or in a civil partnership, in full-time employment and educated to University degree level. The proportions between the *Retrospective and momentary assessments* and the *Retrospective assessments* groups were comparable. Moreover, 43.02% of participants allocated to the *Retrospective plus momentary assessments* group were primigravida, while 56.98% were multigravida. The corresponding figures in the *Retrospective assessments* group were 41.47% and 58.52%, respectively. Furthermore, 94.77% of participants in the *Retrospective plus momentary assessments* group were in their first trimester, and 4.07% in their second; this information was not available for 1.16% of participants. The corresponding figures in the *Retrospective assessments* group were 94.32%, 5.11% and 0.57%, respectively.

Table 24. Participants' socio-demographic characteristics

		Retrospective + Momentary n (% out of 172)	Retrospective n (% out of 176)
Recruitment setting	**General practice**	9 (5.23%)	10 (5.68%)
	Hospital	163 (94.77%)	166 (94.32%)
Age group	**18 – 22 years**	17 (9.88%)	10 (5.68%)
	23 – 27 years	32 (18.60%)	46 (26.14%)
	28 – 32 years	57 (33.14%)	63 (35.80%)
	33 – 37 years	49 (28.49%)	44 (25.00%)
	38+ years	17 (9.88%)	13 (7.39%)
Ethnic background	**White**	146 (84.88%)	153 (86.93%)
	Mixed	3 (1.74%)	1 (0.57%)
	Asian	16 (9.30%)	15 (8.52%)
	Black	0 (0.00%)	0 (0.00%)
	Other ethnic group	6 (3.49%)	7 (3.98%)
	Not disclosed	1 (0.58%)	0 (0.00%)
Relationship status	**Single**	36 (20.93%)	43 (24.43%)
	Married/Civil partnership	126 (73.26%)	116 (65.91%)
	Cohabiting	0 (0.00%)	0 (0.00%)
	Divorced/Widowed/Separated	2 (1.16%)	2 (1.14%)
	Not disclosed	8 (4.65%)	15 (8.52%)
Employment status	**Full-time employment**	93 (54.07%)	82 (46.59%)
	Part-time employment	38 (22.09%)	40 (22.73%)
	Self-employed	16 (9.30%)	22 (12.50%)
	Seeking employment	3 (1.74%)	8 (4.55%)
	Unemployed	22 (12.79%)	24 (13.64%)
	Disability	0 (0.00%)	0 (0.00%)
Highest educational achievement	**Postgraduate degree**	0 (0.00%)	0 (0.00%)
	University degree	93 (54.07%)	100 (56.82%)
	College degree or below	31 (18.02%)	33 (18.75%)
	None	1 (0.58%)	8 (4.55%)
	Other qualification	47 (27.33%)	35 (19.89%)

6.5.1.2. Participants' responses to the Whooley questions and the EPDS

One hundred and forty-eight participants (86.05%) in the *Retrospective plus momentary assessments* group answered *No* to both Whooley questions, 12 (6.98%) answered *Yes* to one question, and 12 (6.98%) answered *Yes* to both questions. In the *Retrospective assessments* group, 144 participants (81.82%) answered *No* to both Whooley questions, 19 (10.80%) answered *Yes* to at least one question, and 13 (7.39%) answered *Yes* to both questions. A chi-squared test of independence showed that there was no association between group allocation and participants' responses to the baseline administration of the Whooley questions, χ^2 (2, N = 348) = 1.6297, p = 0.4427.

The mean baseline EPDS scores in the *Retrospective plus momentary assessments* group was 4.41 (SD = 3.79) and in the *Retrospective assessments* group was 4.42 (SD = 4.00). Overall, 154 women (89.53%) in the *Retrospective plus momentary assessments* group scored between 0 and 9 points on the EPDS, 12 women (6.98%) scored between 10 and 12 points, and six women (3.49%) scored 13 points or more. In the *Retrospective assessments* group, 157 women (89.20%) scored between 0 and 9 points, ten (5.68%) scored between 10 and 12 points, and nine women (5.11%) scored 13 points or more. A chi-squared test of independence showed that there was no association between group allocation and the EPDS scoring intervals for the baseline administration of this scale, χ^2 (2, N = 348) = 0.76488, p = 0.6822.

Concerning question 10 of the EPDS, only one (0.58%) participant in the *Retrospective plus momentary assessments* group and five (2.84%) in the *Retrospective assessments* group scored 1 point or more. A Fisher exact test of independence showed that there was no association between group allocation and risk of self-harm at baseline (as assessed by question ten of the EPDS): p = 0.2149 (95% CI 0.00, 1.82).

6.5.1.3. Proportion of participants who activated BrightSelf after consenting to take part in the study

Of 348 participants who agreed to take part in the study, 251 (72.13%) downloaded, installed and activated BrightSelf onto their handsets and 97 (27.87%) did not (see Table 25). Of those

who activated BrightSelf, 125 (35.92% of the total number of participants recruited into the study) were in the *Retrospective plus momentary assessments* group and 126 (36.21% of the total number of participants recruited into the study) were in the *Retrospective assessments* group. Therefore, 72.67% of participants in the *Retrospective plus momentary assessments* group, and 71.59% of participants in the *Retrospective assessments* group activated BrightSelf.

Table 25. Proportion of participants who did and did not download, install and activate BrightSelf by group allocation

	Downloaded, installed, and activated BrightSelf		
	Yes n (row %)	No n (row %)	**Total**
Retrospective plus momentary	125 (72.67%)	47 (27.33%)	**172**
Retrospective	126 (71.59%)	50 (28.41%)	**176**
Total	**251 (72.13%)**	**97 (27.87%)**	**348**

A chi-squared test of independence was used to assess the relationship between group allocation and activation of BrightSelf. The association between these variables was non-significant, χ^2 (1, N = 348) = 0.011198, p = 0.9157.

Table 26 shows the locality from which the 97 participants who did not download, install and activate BrightSelf were recruited. In the *Retrospective plus momentary assessments* group, the majority of these participants were recruited from antenatal clinics in London (34.04%), whereas in the *Retrospective assessments* group the majority came from antenatal clinics in Shrewsbury (24.00%). In the following localities the number of participants who did not download, install and activate BrightSelf exceeded half of the total number of participants who consented to participate in the study: Burnley (56.25%); Rochdale (50%); Shrewsbury (52.63%); and Telford (66.67%).

Table 26. Proportion of participants who did not activate BrightSelf by group allocation and locality

Locality	Retrospective plus Momentary group n (% out of 47)	Retrospective group n (% out of 50)
Burnley	10 (21.28%)	8 (16.00%)
Burton	1 (2.13%)	0 (0.00%)
Huntingdon	2 (4.26%)	2 (4.00%)
London	16 (34.04%)	10 (20.00%)
Manchester	0 (0.00%)	4 (8.00%)
Rochdale	1 (2.13%)	2 (4.00%)
Shrewsbury	8 (17.02%)	12 (24.00%)
Stoke-on-Trent	3 (6.38%)	6 (12.00%)
Telford	6 (12.77%)	6 (12.00%)
Total	**47**	**50**

6.5.1.4. Types of operating systems

Table 27 shows the number of participants according to the operating system of their smartphones. The majority of participants in both groups owned an iPhone.

Table 27. Overall proportion of operating systems, and by group allocation

Operating System	Retrospective plus Momentary group n (% out of 125)	Retrospective group n (% out of 126)	Total n (% out of 251)
Android	43 (34.4%)	38 (30.16%)	**81 (32.27%)**
iOS	82 (65.6%)	88 (69.84%)	**170 (67.73%)**

6.5.1.5. Overall use of BrightSelf throughout the study

Of the 251 participants who downloaded, installed and activated BrightSelf, 218 (86.85%) completed at least one assessment (either in response to an alert or spontaneously). Of these, 113 were in the *Retrospective plus momentary assessments* group (90.4% of participants who

activated the app in this group) and 105 were in the *Retrospective assessments* group (83.33% of participants who activated the app in this group).

A chi-squared test of independence was used to assess the association between group allocation and the proportion of respondents amongst those who downloaded, installed and activated BrightSelf (see Table 28). The association between these variables was non-significant, χ^2 (1, N = 251) = 2.1602, p = 0.1416.

Table 28. Proportion of respondents by group allocation, amongst those who downloaded, installed and activated BrightSelf

	Respondents n (row %)	**Non-respondents** n (row %)	**Total**
Retrospective plus momentary	113 (90.4%)	12 (9.6%)	**125**
Retrospective	105 (83.33%)	21 (16.67%)	**126**
Total	**218 (86.85%)**	**33 (13.15%)**	**251**

The association between these variables was also non-significant when using the total number of participants randomised into the study: χ^2 (1, N = 348) = 1.1097, p = 0.2921 (see Table 29).

Table 29. Proportion of respondents by group allocation, amongst those randomised into the study

	Respondents n (row %)	**Non-respondents** n (row %)	**Total**
Retrospective plus momentary	113 (65.70%)	59 (34.30%)	**172**
Retrospective	105 (59.66%)	71 (40.34%)	**176**
Total	**218 (62.64%)**	**130 (37.36%)**	**348**

Of 113 respondents in the *Retrospective plus momentary assessments* group, 88 (77.88%) completed at least one EPDS assessment. In this group, the median number of EPDS responses per participant was one (with a minimum of one EPDS response [34 participants] and a maximum of 12 EPDS responses [one participant]), with 283 unique EPDS responses entered. Moreover, 111 participants (98.23%) completed at least one EMA assessment

(which included the contextual questions). The median number of EMA responses per participant was six (with a minimum of one response [11 participants] and a maximum of 61 responses [one participant]), with 1542 unique EMA responses entered. Overall, 86 participants in this group completed both EPDS and EMA assessments, two participants completed only EPDS statements, and 25 participants completed only EMA assessments.

In the *Retrospective assessments* group, all the responses corresponded to EPDS assessment, therefore the total number of non-respondents was 21 participants. The median number of responses per participant was three (with a minimum of one response [29 participants] and a maximum of 32 responses [one participant]), with 462 unique EPDS responses entered.

6.5.2. Adherence to sampling protocol

Three definitions of adherence were considered in this study:

a) Number of participants who completed the total number of expected assessments for the entire duration of the study
b) Number of participants who completed at least one expected assessment in each sampling period for all six sampling periods
c) Number of participants who completed at least one assessment at any point during the study

These were calculated on an intention-to-treat basis.

The third definition prompted the exploration of different patterns of assessment completion.

6.5.2.1. Adherence as the total number of expected assessments

Table 30 shows the number of participants in each group who completed all the expected assessments for each sampling period, regardless of whether or not they completed the assessments in the preceding sampling period. For each sampling period, there was an association between group allocation and adherence rate: χ^2 (1, N = 348), $p < 0.05$ for every sampling period. Therefore, adherence in the *Retrospective assessments* group was higher throughout the study.

Table 30. Number of participants who completed all the expected assessments during each sampling period

	Adherent		Total
	Yes n (row %)	No n (row %)	
Sampling period: 1			
Retrospective plus momentary	2 (1.16%)	170 (98.84%)	**172**
Retrospective	81 (46.02%)	95 (53.98%)	**176**
Sampling period: 2			
Retrospective plus momentary	0 (0.00%)	172 (100%)	**172**
Retrospective	60 (34.09%)	116 (65.91%)	**176**
Sampling period: 3			
Retrospective plus momentary	1 (0.58%)	171 (99.42%)	**172**
Retrospective	46 (26.14%)	130 (73.86%)	**176**
Sampling period: 4			
Retrospective plus momentary	0 (0.00%)	172 (100%)	**172**
Retrospective	40 (22.73%)	136 (77.27%)	**176**
Sampling period: 5			
Retrospective plus momentary	0 (0.00%)	172 (100%)	**172**
Retrospective	29 (16.48%)	147 (83.52%)	**176**
Sampling period: 6			
Retrospective plus momentary	0 (0.00%)	172 (100%)	**172**
Retrospective	22 (12.50%)	154 (87.50%)	**176**

Figure 14 shows the levels of adherence for each sampling period. Overall, the number of participants in the *Retrospective assessments* group who adhered to the sampling protocol declined over time. A temporal trend for the *Retrospective plus momentary assessments*

group could not be determined since not enough participants completed all the expected assessments in each sampling period.

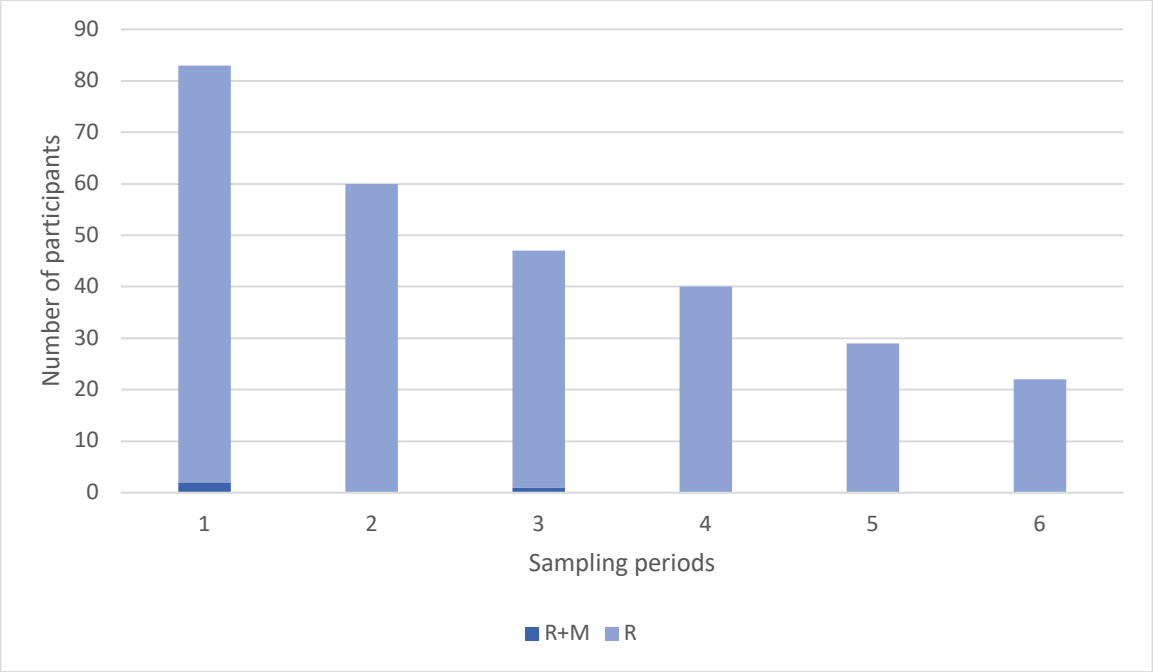


Figure 14. Number of participants who completed all the expected assessments during each sampling period.

Table 31 shows the adherence rate for each sampling period, dependent on participants having completed the immediately preceding sampling period. There was a significant association between group allocation and adherence in the first sampling period: χ^2 (1, N = 348) = 93.932, p = 2.2 x 10^{-16}. A Fisher exact test of independence revealed that there was no association between these variables in the second sampling period: p = 0.2177 (95% CI 0.00, 4.69). The remaining four sampling periods could not be compared as the adherence rate in the *Retrospective plus momentary assessments* group was zero for each one of them. Adherence rate in the *Retrospective assessments* group increased gradually until the fourth sampling period, after which it declined. However, it is worth noting that the denominator used in these calculations kept decreasing with each successive sampling period.

Table 31.Number of participants who completed all the expected assessments during each sampling period, after having completed the assessments in the immediately preceding sampling periods

	Adherent		**Total**
	Yes n (row %)	No n (row %)	
Sampling period: 1			
Retrospective plus momentary	2 (1.16%)	170 (98.84%)	**172**
Retrospective	81 (46.02%)	95 (53.98%)	**176**
Sampling period: 2 (after having completed sampling period 1)			
Retrospective plus momentary	0 (0.00%)	2 (100%)	**2**
Retrospective	44 (54.32%)	37 (45.68%)	**81**
Sampling period: 3 (after having completed sampling periods 1 and 2)			
Retrospective plus momentary	0	0	**0**
Retrospective	28 (63.64%)	(36.36%)	**44**
Sampling period: 4 (after having completed sampling periods 1, 2 and 3)			
Retrospective plus momentary	0	0	**0**
Retrospective	17 (60.71%)	11 (39.29%)	**28**
Sampling period: 5 (after having completed sampling periods 1, 2, 3 and 4)			
Retrospective plus momentary	0	0	**0**
Retrospective	10 (58.82%)	7 (41.18%)	**17**
Sampling period: 6 (after having completed sampling periods 1, 2, 3, 4 and 5)			
Retrospective plus momentary	0	0	**0**
Retrospective	3 (30.00%)	7 (70.00%)	**10**

Table 32 shows the number of participants who completed all the expected assessments for the entire duration of the study (i.e., all the expected assessments for each sampling period for all six sampling periods), as a proportion of the number of participants who were randomised into the study. In the *Retrospective plus momentary assessments* group, each participant was expected to complete 12 EPDS and 72 EMA assessments over six months (two EPDS assessments and 12 EMA assessments each month). In this group, none of the participants completed all the expected assessments; therefore, the adherence rate was 0%. In the *Retrospective assessments* group, each participant was expected to complete six EPDS assessments over six months. Three participants completed all the assessments, resulting in an adherence rate of 1.70%. The overall response rate in the study was 0.86%.

Table 32. Adherence rate by allocation group, defined as the number of participants who completed all the expected assessments as a proportion of the number of participants randomised into the study

	Adherence		
	Yes n (row %)	No n (row%)	**Total**
Retrospective plus momentary	0 (0.00%)	172 (100%)	**172**
Retrospective	3 (1.70%)	173 (98.30%)	**176**
Total	**3 (0.86%)**	**345 (99.14%)**	**348**

A Fisher exact test of independence was used to assess the association between group allocation and this definition of adherence. The association between these variables was non-significant: p = 0.2479 (95% CI 0.00, 2.47).

6.5.2.2. Adherence as at least one expected assessment in each sampling period for all sampling periods

This definition only affected how the data for participants in the *Retrospective and momentary assessments* group were interpreted, as participants in this group were expected to complete two EPDS assessments and 12 EMA assessments during each sampling period (unlike participants in the *Retrospective* group who were expected to complete only one EPDS assessment during each sampling period). Therefore, any participant completing at least one of the 14 expected assessments was considered adherent.

Table 33 shows the number of participants who completed at least one expected assessment during each sampling period, regardless of whether or not they completed the preceding sampling period. Group allocation and adherence were associated in the first sampling period: χ^2 (1, N = 348) = 9.8744, p = 0.001676, showing that adherence in the *Retrospective plus momentary assessments* group was higher than in the *Retrospective assessments* group. There was no association between these two variables in the remaining sampling periods.

Table 33. Number of participants who completed at least one expected assessment during each sampling period, regardless of whether they completed the preceding sampling periods

	Adherent		**Total**
	Yes n (row %)	No n (row %)	
Sampling period: 1			
Retrospective plus momentary	109 (63.37%)	63 (36.63%)	**172**
Retrospective	81 (46.02%)	95 (53.98%)	**176**
Sampling period: 2			
Retrospective plus momentary	67 (38.95%)	105 (61.05%)	**172**
Retrospective	60 (34.09%)	116 (65.91%)	**176**
Sampling period: 3			
Retrospective plus momentary	62 (36.05%)	110 (63.95%)	**172**
Retrospective	46 (26.14%)	130 (73.86%)	**176**
Sampling period: 4			
Retrospective plus momentary	48 (27.91%)	124 (72.09%)	**172**
Retrospective	40 (22.73%)	136 (77.27%)	**176**
Sampling period: 5			
Retrospective plus momentary	37 (21.51%)	135 (78.49%)	**172**
Retrospective	29 (16.48%)	147 (83.52%)	**176**
Sampling period: 6			
Retrospective plus momentary	32 (18.60%)	140 (81.40%)	**172**
Retrospective	22 (12.50%)	154 (87.50%)	**176**

Figure 15 shows that for both groups the number of participants adhering to the sampling protocol declined with each successive sampling period.

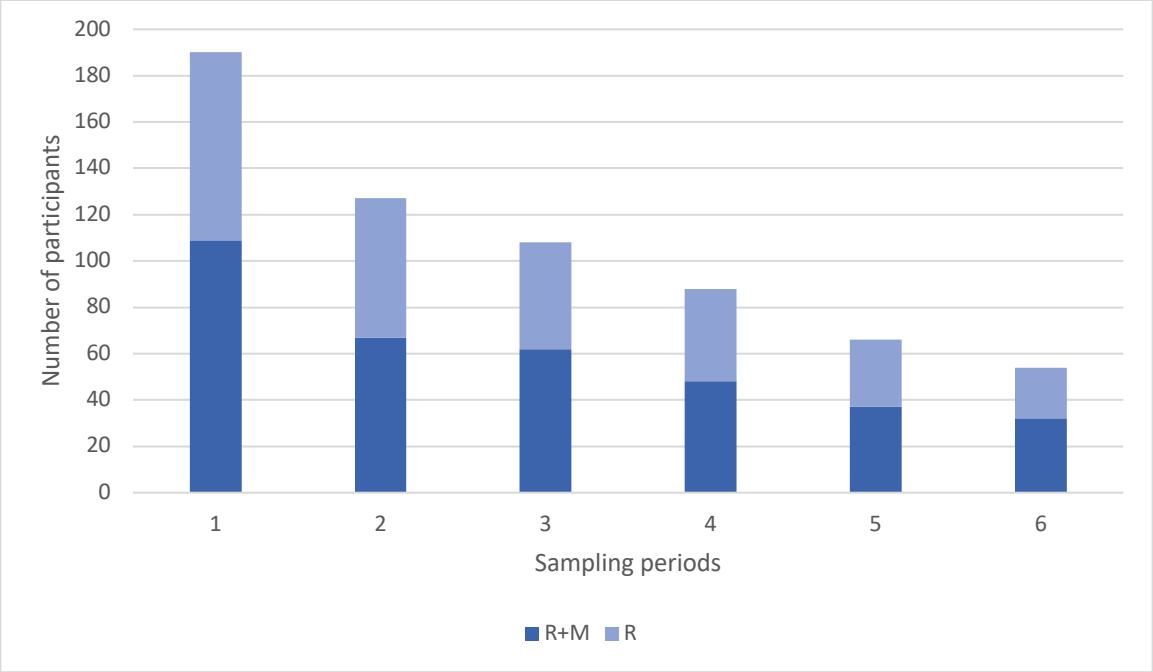


Figure 15. Number of participants who completed at least one expected assessment during each sampling period

Table 34 shows the number of participants who completed at least one expected assessment in each sampling period, after having completed at least one expected assessment in the immediately preceding sampling period. A chi-squared test of independence showed that there was an association between group allocation and adherence in the first sampling period: χ^2 (1, N = 348) = 9.8744, p = 0.001676. Similarly, a Fisher exact test of independence showed that there was an association between group allocation and adherence in the sixth sampling period: p = 0.01906 (95% CI 1.19, 53.00). There was no association between these variables in the other sampling periods.

Table 34. Number of participants who completed at least one expected assessment in each sampling period, dependent on having completed the preceding sample periods

	Adherent		Total
	Yes n (row %)	No n (row %)	
Sampling period: 1			
Retrospective plus momentary	109 (63.37%)	63 (36.63%)	**172**
Retrospective	81 (46.02%)	95 (53.98%)	**176**
Sampling period: 2 (after having completed sampling period 1)			
Retrospective plus momentary	66 (60.55%)	43 (39.45%)	**109**
Retrospective	44 (54.32%)	37 (39.51%)	**81**
Sampling period: 3 (after having completed sampling periods 1 and 2)			
Retrospective plus momentary	53 (80.30%)	13 (19.70%)	**66**
Retrospective	28 (63.64%%)	16 (36.36%)	**44**
Sampling period: 4 (after having completed sampling periods 1, 2 and 3)			
Retrospective plus momentary	40 (75.47%)	13 (24.53%)	**53**
Retrospective	17 (60.71%)	11 (39.29%)	**28**
Sampling period: 5 (after having completed sampling periods 1, 2, 3 and 4)			
Retrospective plus momentary	29 (72.50%)	11 (27.50%)	**40**
Retrospective	10 (58.82%)	7 (41.18%)	**17**
Sampling period: 6 (after having completed sampling periods 1, 2, 3, 4 and 5)			
Retrospective plus momentary	22 (75.86%)	7 (24.14%)	**29**
Retrospective	3 (30.00%)	7 (70.00%)	**10**

Table 35 shows the number of participants who completed at least one expected assessment during each sampling period for all six sampling periods. With this definition, the adherence rates in the *Retrospective plus momentary assessments* and in the *Retrospective assessments* groups were 12.79% and 1.70%, respectively. The overall adherence rate was 5.56%.

Table 35. Adherence rate by group allocation for the whole duration of the study, defined as the number of participants who completed at least one of the expected assessments in each sampling period for all sampling periods as a proportion of the number of participants randomised into the study

	Adherence		
	Yes n (row %)	No n (row %)	**Total**
Retrospective plus momentary	22 (12.79%)	150 (87.21%)	**172**
Retrospective	3 (1.70%)	173 (98.30%)	**176**
Total	**25 (5.56%)**	**323 (94.44%)**	**348**

A chi-squared test of independence showed a significant association between group allocation and this definition of adherence: χ^2 (1, N = 348) = 14.414, p = 0.0001467. However, it is worth noting that of the 22 participants in the *Retrospective plus momentary assessments* group who were considered adherent under this definition, only five completed at least one EPDS report in each sampling period.

6.5.2.3. Adherence as the number of participants who completed at least one assessment at any point during the study

Table 36 shows the number of participants in each group who completed at least one expected assessment at any point during the study, as a proportion of the total number of participants randomised into the study. Defined as such, the adherence rates in the *Retrospective plus momentary assessments* and in the *Retrospective assessments* groups were 63.95% and 56.82%, respectively. The overall response rate was 60.34%.

Table 36. Adherence rate by group allocation, defined as the number of participants who completed at least one expected assessment as a proportion of the number of participants randomised into the study

	Adherence		Total
	Yes n (row %)	**No** n (row %)	
Retrospective plus momentary	110 (63.95%)	62 (36.05%)	**172**
Retrospective	100 (56.82%)	76 (43.18%)	**176**
Total	**210 (60.34%)**	**138 (39.66%)**	**348**

A chi-squared test of independence was used to assess the association between group allocation and this definition of adherence. The association between these variables was non-significant: χ^2 (1, N = 348) = 1.5646, p = 0.211.

6.5.2.4. Patterns of assessment completion

Table 37 shows the number of participants according to the number of assessment periods they completed. Only EPDS assessments could be considered for this between group comparison. Assessment periods consisted of six one-month intervals, calculated from the date of the first BrightSelf notification (i.e., two days after activating the app). The majority of participants did not complete any assessments and approximately a quarter of them completed one assessment period. The remaining groups accounted for less than ten percent of all participants, with less than two percent of participants completing all six assessment periods. The association between group allocation and the number of assessment periods that were completed was evaluated using a chi-squared test of independence with Monte Carlo simulation (as more than 20% of the cells had expected frequencies of less than five), χ^2 (NA, N = 349) = 17.276, p = 0.006997. Since the association between these variables was significant, the adjusted standardised residuals were included in Table 37. A visual inspection of these residuals revealed that the observed frequency of participants in the *Retrospective assessments* group who completed three assessment periods exceeded the expected frequency by more than the critical value of 1.96 for a significance level of 0.05.

Table 37. Number of participants according to the number of assessment periods they completed (EPDS assessments only)

		Number of assessment periods completed - n (row %)							Total
		0	**1**	**2**	**3**	**4**	**5**	**6**	
Retrospective plus momentary	Observed	86 (50.00%)	50 (29.07%)	14 (8.14%)	7 (4.07%)	10 (5.81%)	4 (2.33%)	1 (0.58%)	**172**
	Expected	77.6	43	14.83	15.82	13.84	4.94	1.98	
	Standardised residuals	1.81	1.73	-0.32	-3.27	-1.51	-0.60	-0.98	
Retrospective	Observed	71 (40.34%)	37 (21.02%)	16 (9.10%)	25 (14.20%)	18 (10.23%)	6 (3.41%)	3 (1.70%)	**176**
	Expected	79.4	44	15.17	16.18	14.16	5.06	2.02	
	Standardised residuals	-1.81	-1.73	0.32	3.27	1.51	0.60	0.98	
Total		**157 (45.11%)**	**87 (25.00%)**	**30 (8.62%)**	**32 (9.20%)**	**28 (8.05%)**	**10 (2.87%)**	**4 (1.15%)**	**348**

Table 38 shows the number of participants who completed one assessment period according to the assessment period they completed. The majority of participants in this category completed the first assessment period, and approximately ten percent completed the second assessment period. Only three participants completed their only assessment period half way through the study (i.e., during assessment periods three, four or five). No participant completed her only assessment towards the end of the study.

Table 38. Number of participants who completed one assessment period according to the assessment period they completed (EPDS assessments only)

	Assessment period that was completed – n (row %)						Total
	1	**2**	**3**	**4**	**5**	**6**	
Retrospective plus momentary	43 (86.00%)	5 (10.00%)	1 (2.00%)	0 (0.00%)	1 (2.00%)	0 (0.00%)	**50**
Retrospective	32 (86.49%)	4 (10.81%)	0 (0.00%)	1 (2.70%)	0 (0.00%)	0 (0.00%)	**37**
Total	**75 (86.21%)**	**9 (10.34%)**	**1 (1.15%)**	**1 (1.15%)**	**1 (1.15%)**	**0 (0.00%)**	**87**

Tables 39 to 42 present the number of participants who completed between two and five assessment periods according to the assessment periods they completed. The majority of participants tended to complete the initial assessment periods (usually a combination of the

first three assessment periods). Participants who completed three or more assessment periods showed a tendency to do so in a consecutive fashion (for example, assessment periods 1, 2, 3 and 4).

Table 39. Number of participants who completed two assessment periods according to the assessment periods they completed (EPDS assessments only)

	Assessment periods that were completed – n (row %)							**Total**
	1 and 2	1 and 3	1 and 4	1 and 5	1 and 6	2 and 3	3 and 5	
Retrospective plus momentary	6 (42.86%)	5 (35.71%)	0 (0.00%)	2 (14.29%)	0 (0.00%)	0 (0.00%)	1 (7.14%)	**14**
Retrospective	5 (31.25%)	3 (18.75%)	3 (18.75%)	0 (0.00%)	1 (6.25%)	4 (25.00%)	0 (0.00%)	**16**
Total	**11 (36.67%)**	**8 (26.67%)**	**3 (10.00%)**	**2 (6.67%)**	**1 (3.33%)**	**4 (13.33%)**	**1 (3.33%)**	**30**

Table 40. Number of participants who completed three assessment periods according to the assessment periods they completed (EPDS assessments only)

	Assessment periods that were completed – n (row %)									**Total**
	1, 2 and 3	1, 2 and 4	1, 2 and 5	1, 2 and 6	1, 3 and 4	1, 3 and 5	1, 3 and 6	1, 4 and 6	2, 3 and 4	
Retrospective plus momentary	4 (57.14%)	1 (14.29%)	0 (0.00%)	0 (0.00%)	2 (28.57%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	**7**
Retrospective	8 (32.00%)	1 (4.00%)	4 (16.00%)	1 (4.00%)	5 (20.00%)	3 (12.00%)	1 (4.00%)	1 (4.00%)	1 (4.00%)	**25**
Total	**12 (37.50%)**	**2 (6.25%)**	**4 (12.50%)**	**1 (3.13%)**	**7 (21.88%)**	**3 (9.38%)**	**1 (3.13%)**	**1 (3.13%)**	**1 (3.13%)**	**32**

Table 41. Number of participants who completed four assessment periods according to the assessment periods they completed (EPDS assessments only)

	Assessment periods that were completed – n (row %)						
	1, 2, 3 and 4	1, 2, 3 and 5	1, 2, 3 and 6	1, 2, 4 and 5	1, 3, 4 and 5	2, 3, 4 and 5	**Total**
Retrospective plus momentary	2 (20.00%)	4 (40.00%)	1 (10.00%)	2 (20.00%)	1 (10.00%)	0 (0.00%)	**10**
Retrospective	9 (50.00%)	4 (22.22%)	1 (5.56%)	2 (11.11%)	0 (0.00%)	2 (11.11%)	**18**
Total	**11 (39.29%)**	**8 (28.57%)**	**2 (7.14%)**	**4 (14.29%)**	**1 (3.57%)**	**2 (7.14%)**	**28**

Table 42. Number of participants who completed five assessment periods according to the assessment periods they completed (EPDS assessments only)

	Assessment periods that were completed – n (row %)					**Total**
	1, 2, 3, 4 and 5	1, 2, 3, 4 and 6	1, 2, 4, 5 and 6	1, 3, 4, 5 and 6	2, 3, 4, 5 and 6	
Retrospective plus momentary	2 (50.00%)	1 (25.00%)	0 (0.00%)	0 (0.00%)	1 (25.00%)	**4**
Retrospective	3 (50.00%)	1 (16.67%)	1 (16.67%)	1 (16.67%)	0 (0.00%)	**6**
Total	**5 (50.00%)**	**2 (20.00%)**	**1 (10.00%)**	**1 (10.00%)**	**1 (10.00%)**	**10**

6.5.2.5. Drop-out rates

Table 43 shows the number of participants in each experimental group who dropped out of the study. This outcome was defined in relation to the number of participants who completed the six-month participation period as a proportion of the total number of participants randomised into the study. The overall drop-out rate was 84.48%, whilst the drop-out rates for the *Retrospective and momentary assessments* and the *Retrospective assessments* groups were 81.40% and 87.50%, respectively.

Table 43. Drop-out rates by group allocation

	Drop-out		Total
	Yes n (row %)	**No** n (row %)	
Retrospective plus momentary	140 (81.40%)	32 (18.60%)	**172**
Retrospective	154 (87.50%)	22 (12.50%)	**176**
Total	**294 (84.48%)**	**54 (15.52%)**	**348**

A chi-squared test of independence showed that there was no association between drop-out rates and group allocation: χ^2 (1, N = 348) = 2.0291, p = 0.1543.

6.5.2.6. EPDS and EMA assessments

6.5.2.6.1. EPDS scores between the Retrospective and Momentary, and the Retrospective groups

The intention was to use a linear mixed effects model to evaluate the effect of assessment strategy (i.e., *Retrospective plus momentary assessments* or *Retrospective assessments*) and of within-participant differences over time on EPDS scores [9-11]. By implementing this model, only the EPDS scores at the end of the sampling period in the *Retrospective and momentary assessments* group would have been considered as it overlapped with the momentary assessments (which represented the key difference between the two experimental groups).

Fitting a linear mixed effects model requires for each observation to be treated individually in the dataset. Therefore, with 348 participants, there were 2088 expected observations (i.e., six sampling periods x 348 participants). However, in the original dataset, there were only 313 observations (approximately 15% of the expected observations). In other words, the percentage of missing data was 85%. Moreover, there was a significant relationship between missing data on the EPDS scores and all the variables collected during the baseline assessment, with the exception of marital status, obstetric history (both gravidity and parity), pregnancy trimester, and previous depression. This suggests that data were missing not at random.

Maximum likelihood estimation or multiple imputation methods were not used to deal with missing data, given the large volume of missing data and because the missing-at-random assumption required to implement them was violated [12]. Implementing these methods could have led to biased estimates of the parameters of the linear mixed effects model. Moreover, sensitivity analyses were not conducted as they require a thorough understanding of the mechanisms for missing data and the current design was not appropriate to identify these.

6.5.2.6.2. Effect of momentary assessments on EPDS scores

The intention was to use a linear mixed effects model to explore the effect of the momentary question on the EPDS scores at each tail-end of the assessment periods in the *Retrospective and momentary assessments* group. As with the previous analysis, there was a large volume of missing data. Of 1032 expected EPDS observations (i.e., 172 participants allocated to this group x six final EPDS assessments), only 68 were obtained. Therefore, the percentage of missing data was 93.41%. Maximum likelihood estimation or multiple imputation methods were not used to fit a linear mixed effects model for the same reasons outlined previously.

6.5.2.7. Alerts

Overall, 51 participants (25 in the *Retrospective plus momentary assessments* group and 26 in the *Retrospective assessments* group) triggered alerts during the study, accounting for 88 unique alerts. This number of participants accounted for 14.66% of the number of participants randomised into the study, which is compatible with the expected risk of depression in this population.

Table 44 shows the number and type of alerts that were triggered. Orange alerts were the most common type, accounting for 46.59% of all alerts. Red alerts accounted for 34.09% of alerts, and yellow alerts for 19.32%. Moreover, the majority of alerts (57.95%) came from participants in the *Retrospective assessments* group. A chi-squared test of independence showed that there was an association between group allocation and type of alert: χ^2 (2, N = 88) = 15.799, p = 0.0004.

Table 44.Number of alerts by type and group allocation

	Yellow alerts n (row %)	Orange alerts n (row %)	Red alerts n (row %)	Total
Retrospective plus momentary	14 (37.84%)	16 (43.24%)	7 (18.92%)	**37**
Retrospective	3 (5.88%)	25 (49.02%)	23 (45.10%)	**51**
Total	**17 (19.32%)**	**41 (46.59%)**	**30 (34.09%)**	**88**

Table 45 shows the number of participants in each experimental group according to the number of alerts that they triggered. The majority of participants (68.63%) triggered one alert: 80% in the *Retrospective plus momentary assessments* group and 57.69% in the *Retrospective assessments* group. The two second largest categories were participants who triggered two alerts and participants who triggered four alerts, with 11.76% each.

Table 45. Number of participants in each experimental group according to the number of alerts the triggered

	One alert n (row %)	Two alerts n (row %)	Three alerts n (row %)	Four alerts n (row %)	Five alerts n (row %)	Six alerts n (row %)	Total
Retrospective plus momentary	20 (80.00%)	2 (8.00%)	0 (0.00%)	2 (8.00%)	1 (4.00%)	0 (0.00%)	**25**
Retrospective	15 (57.69%)	4 (15.38%)	2 (7.69%)	4 (15.38%)	0 (0.00%)	1 (3.85%)	**26**
Total	**35** **(68.63%)**	**6** **(11.76%)**	**2** **(3.92%)**	**6** **(11.76%)**	**1** **(1.96%)**	**1** **(1.96%)**	**51**

A chi-squared test of independence was used to assess the relationship between group allocation and the number of alerts triggered. No association was observed between these variables: χ^2 (5, N = 51) = 6.030, p = 0.3033.

6.5.2.8. Acceptance

Overall, 62 participants (25 in the *Retrospective plus momentary* group and 37 in the *Retrospective* group) completed the post-study acceptance survey. This accounted for 29.52% of participants who completed at least one assessment throughout the study.

6.5.2.8.1. Ease of use

Figure 16 shows the number of participants agreeing or disagreeing with the statement *The app was easy to use*. Overall, 47 participants (75.81%) strongly agreed with this statement, and seven participants (11.29%) agreed. Only two participants (3.23%) disagreed or strongly disagreed with this statement.

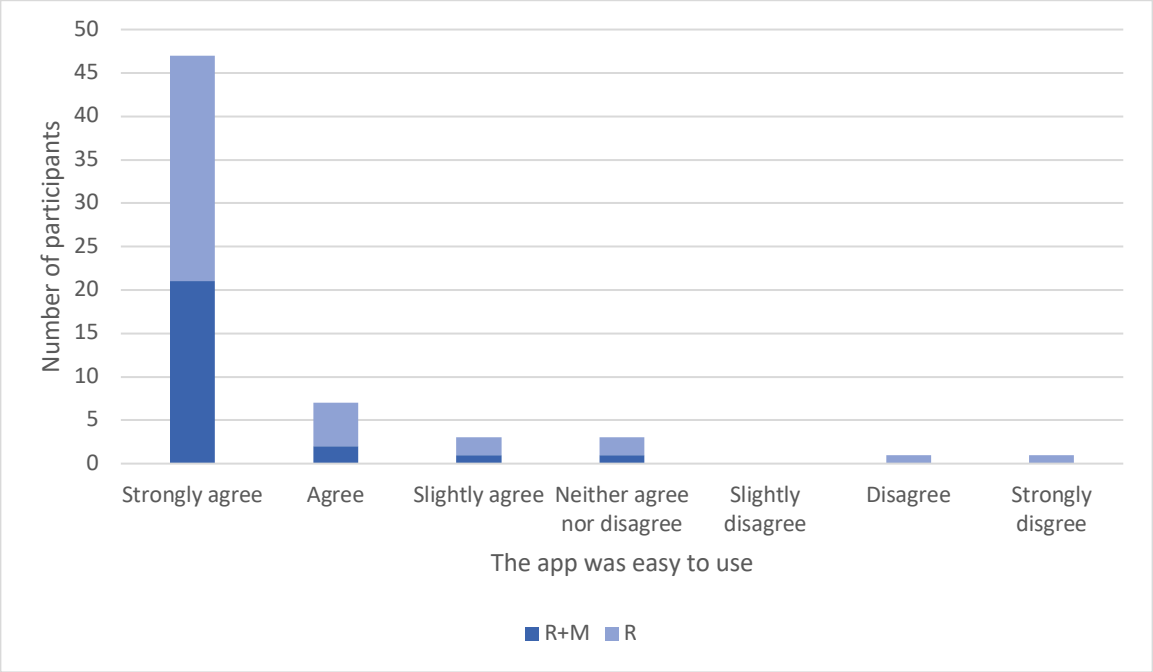


Figure 16. Number of participants agreeing/disagreeing with the statement: the app was easy to use

6.5.2.8.2. Engagement

Figure 17 shows the number of participants agreeing or disagreeing with the statement *I found using this app an engaging experience*.

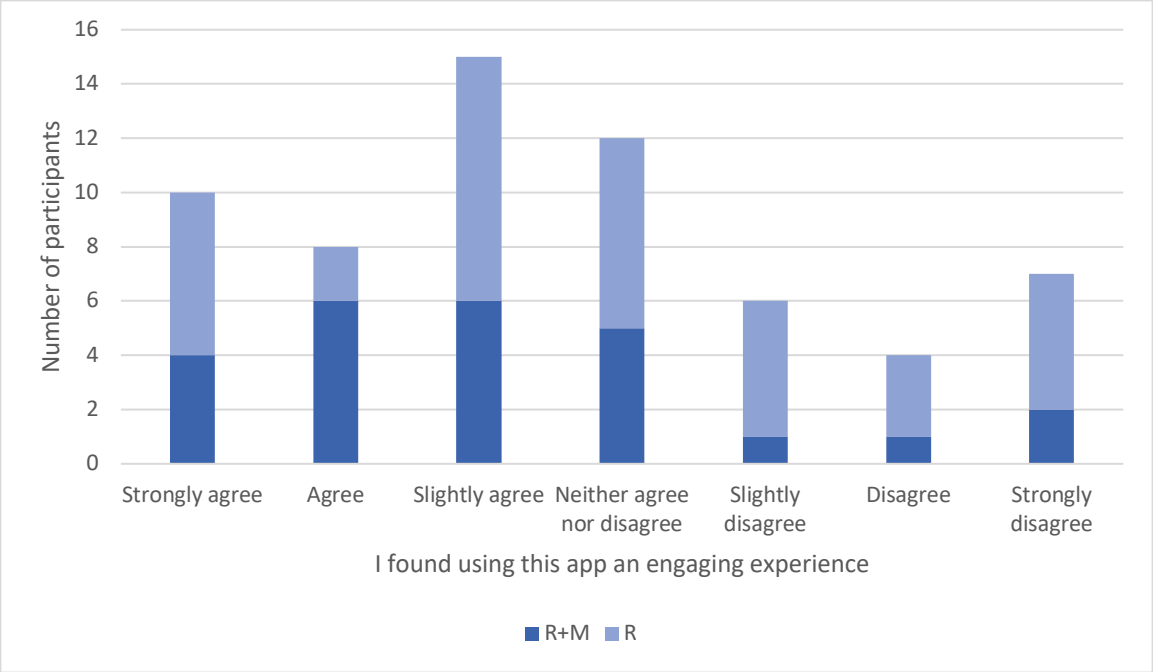


Figure 17. Number of participants agreeing/disagreeing with the statement: I found using this app an engaging experience

Ten participants (16.13%) strongly agreed with this statement, whereas seven (11.29%) strongly disagreed with it. Overall, the majority of participants (24.19%) slightly agreed with the statement presented in this question.

6.5.2.8.3. Perceived usefulness of the assessments

Figure 18 shows the number of participants agreeing or disagreeing with the statement *I found the assessments useful*. Overall, 14 participants (22.58%) strongly agreed with this statement, nine (14.52%) agreed and 17 (27.42%) slightly agreed. Five participants (8.06%) strongly disagreed with this statement.

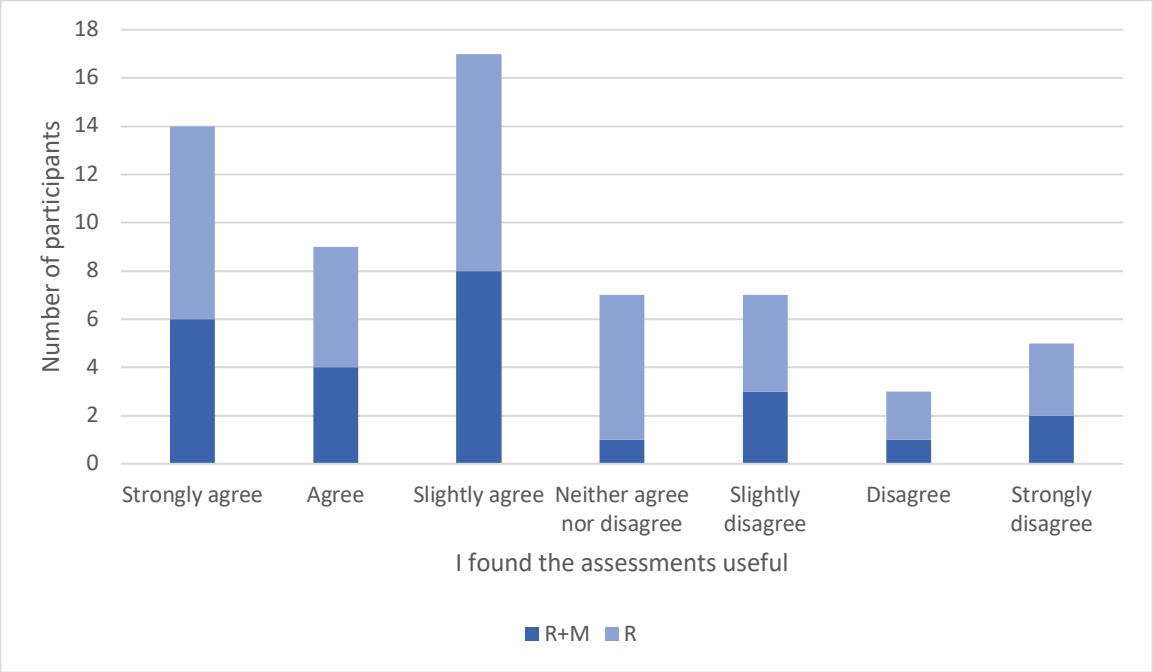


Figure 18. Number of participants agreeing/disagreeing with the statement: I found the assessments useful

6.5.2.8.4. Frequency of reporting

Figure 19 shows the number of participants agreeing or disagreeing with the statement *I was asked to provide reports…* (response options were related to the frequency of reporting). Overall, 30 participants (48.39%) felt that the frequency of assessments was *the right amount* and 20 participants (32.26%) felt that the assessments were too infrequent. However, of these 20 participants, only three (15%) were in the *Retrospective plus momentary* group. In addition, six participants (9.68%) felt that the assessments were too frequent, and three (4.84%) felt that the assessments were much too frequent.

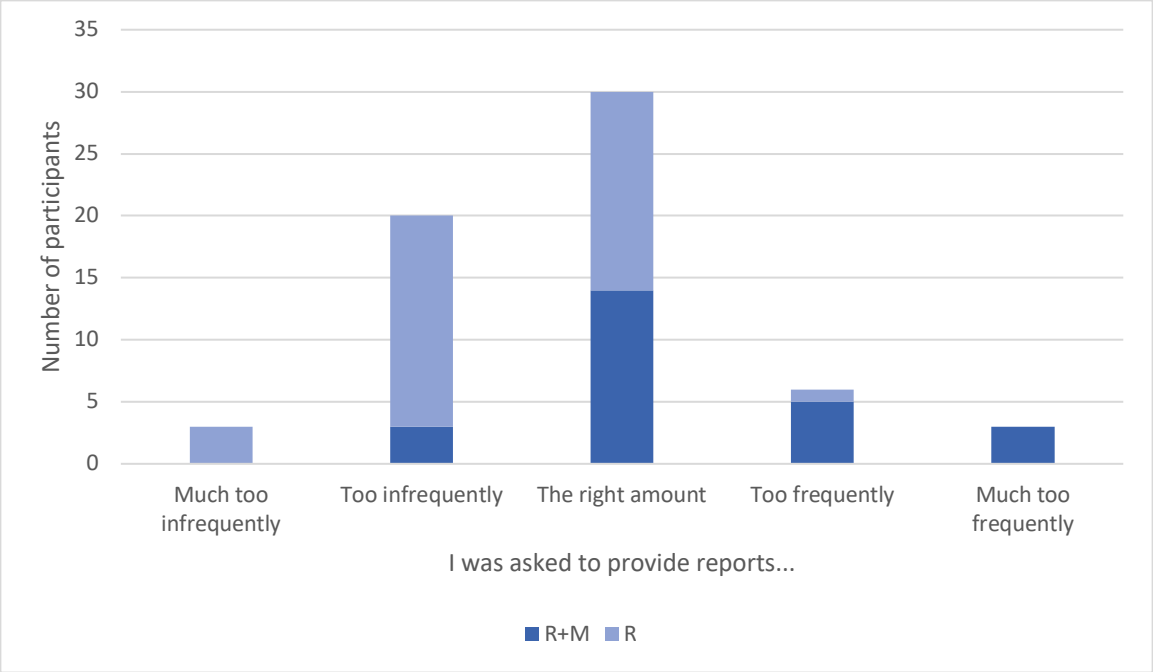


Figure 19. Number of participants agreeing/disagreeing with the statement: I was asked to provide reports… (related to their frequency)

6.5.2.8.5. Perceived needs

Figure 20 shows the number of participants agreeing or disagreeing with the statement *The experience of using this app met my needs*. Overall, 17 participants (27.42%) strongly agreed with this statement, with 11 participants (64.71%) being in the *Retrospective* group. In addition, 16 participants (25.81%) neither agreed nor disagreed with this statement. Moreover, 11 participants (17.74%) were on the disagree side of this scale.

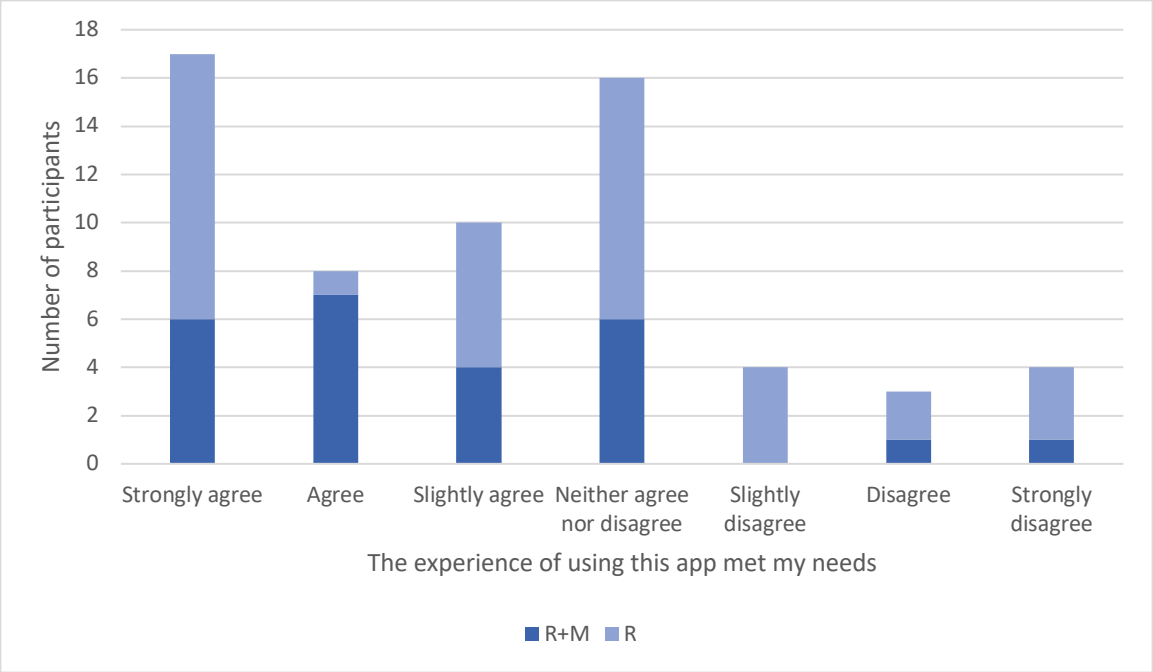


Figure 20. Number of participants agreeing/disagreeing with the statement: the experience of using this app met my needs

6.5.2.8.6. Perceived usefulness

This question was only given to those participants allocated to the *Retrospective plus momentary assessments* group, and presented as the statement *I found it useful to compare reports made right now with those made over the past 7 days*.

Figure 21 shows the number of participants agreeing or disagreeing with this statement. Overall, five participants (20%) agreed with this statement. Another 20% of participants said that they neither agreed nor disagreed with the statement. Moreover, four participants (16%) strongly agreed with this statement, and another four participants (16%) slightly agreed.

6.5.2.8.7. Reasons for using BrightSelf

Of those participants who answered the question *What were your motivations in using this app?*, 31 (50%) declared that it was to support research. In addition, 20 participants expressed their desire to help others who had experienced mental health problems or declared that their main motivation was due to their own personal experience of mental health problems during pregnancy.

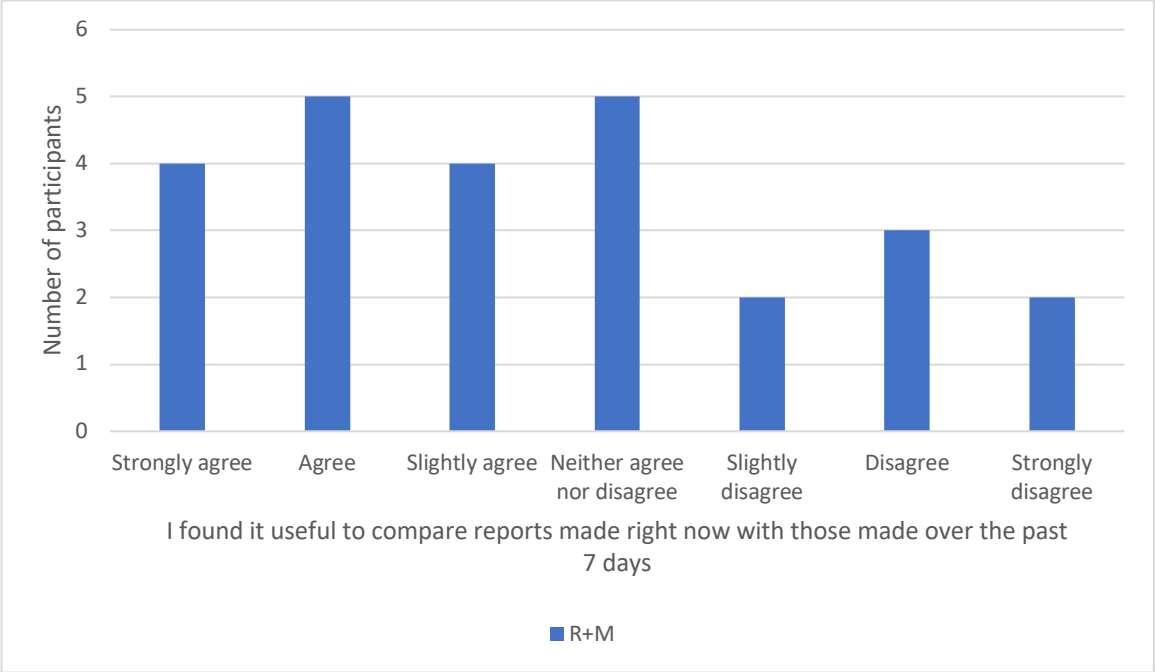


Figure 21. Number of participants in the Retrospective plus momentary group agreeing/disagreeing with the statement: I found it useful to compare reports made right now and those made over the past 7 days

6.6. Discussion

This study evaluated the feasibility of using a bespoke app, BrightSelf, running on participants' own devices for the assessment and monitoring of risk of depression during the last two trimesters of pregnancy. Its primary focus was on patient adherence as this is considered to be essential for data completeness and for the effectiveness of digital health interventions [1]. Particularly, when unsupervised, longitudinal data collection in naturalistic settings is required. This study also explored implementation challenges associated with the deployment of BrightSelf in antenatal clinics.

Optimising adherence is considered a critical design factor of digital health interventions. To achieve this, an intervention ought to be sensitive to and support the needs, motivation and goals of its users [1]. That was the aim of the PPI sessions that informed BrightSelf and the study protocol [4]. Pregnant women who participated in these sessions expressed that their motivation for using such a system would be to facilitate self-reflection, and to enable comparisons with normative experiences of mood changes during pregnancy. This led to the incorporation of momentary questions in the *Retrospective plus momentary assessments* group, to assess if they could satisfy these needs and lead to better adherence with BrightSelf.

Hence, promoting the completion of the retrospective assessments that clinicians required. The findings however, did not support this hypothesis.

6.6.1. Participant adherence

Adherence was explored through an analysis of BrightSelf usage, adherence rates, drop-out rates, participants' answers to a post-study acceptance survey, and through the levels of missing data identified whilst attempting to fit a linear mixed effects model. Overall, the results highlight the complex nature of adherence and the implications of this complexity for the development and implementation of similar mHealth systems.

With regards to usage, it would appear that participants found the idea of BrightSelf and the associated sampling protocols acceptable. Indeed, 72% of participants downloaded, installed and activated BrightSelf. Moreover, 87% of installers went on to complete at least one assessment. This is particularly compelling as participants were willing to install what was essentially a health system onto their personal devices, indicating that there might be a desire for such a system in the antenatal period.

Other findings however, suggest that the intensity of the sampling protocols might have been more demanding than initially thought. For example, of the 14784 entries expected in the *Retrospective plus momentary assessments* group, only 1825 were submitted (approximately 12%). Similarly, only 462 out of 1032 expected entries were submitted by participants in the *Retrospective assessments* group (approximately 45%). As a result of these high levels of missing data, and the association between missing values and baseline variables, the effect of assessment strategy on EPDS scores could not be modelled. Furthermore, there was a high overall drop-out rate (approximately 84% of participants).

Adherence rates helped to complement the picture provided by these measure of app usage. Adherence can encompass different conceptual frameworks that range from traditional notions in which patients are expected to comply with pre-specified requests, and current notions that acknowledge patients' active role in making decisions about their own healthcare [5, 13]. This can create a tension when the conceptual framework that informs

expected levels of adherence is not aligned with the way in which patients interact with the technology [1, 13-15].

This tension was explored through the differences between the following operational definitions of adherence rate: (i) participants who completed all the expected assessments for the entire duration of the study; (ii) participants who completed at least one expected assessment during each sampling period; and (iii) participants who completed at least one expected assessment at any point during the study. These operational definitions were thought to represent the continuum of conceptual frameworks of adherence, the first definition being compatible with traditional notions and the third being compatible with current notions, with the second definition representing a middle point.

The findings in relation to these definitions suggest that, in naturalistic settings, patients use health apps in a way that suits their individual needs, rather than as prescribed by healthcare practitioners. In line with other findings in the literature [1, 14, 15], expecting participants to complete all the sampling protocols or to complete all the expected assessments led to extremely poor adherence rates. These rates increased as the definition departed from traditional, prescriptive notions of adherence. This increase however, was driven by the momentary assessments and not by the retrospective assessments required by midwives.

These findings also suggest that adherence might be better understood as a conglomerate of multiple patterns of utilisation rather than as a summative process [6]. Indeed, several patterns were identified in this study, which suggests that strategies to promote adherence require better targeting, perhaps at the individual level. Moreover, completing an assessment appeared to increase the likelihood of completing subsequent ones. This suggests that strategies to promote adherence might also need to be adaptive. Of course, more effort is required to understand the factors that can influence assessment completion, and how these factors vary across individuals and contexts. Since adherence was at its highest during the first half of the study, and it tended to decline over time, additional resources (e.g., direct messages from a member of the clinical care team) might be required at the start of the data collection process in order to encourage longer-term adherence.

Several explanations could account for these changes in adherence rates. One possibility is that the benefits (perceived or real) of the assessments did not outweigh the commitment required to complete them [1], or that they were not able to support participants' goals. Messages displayed after completing any type of assessment were too generic, were not responsive to the data and provided no clear actions for maintaining or improving mental health (other than contacting a GP or midwife if the recommended EPDS thresholds were exceeded). Furthermore, visual representations of the data might not have encouraged or facilitated self-reflection [13]. For example, they did not plot against normative mood changes, which was one of the factors identified during the PPI sessions [4]. In addition, they did not offer the possibility to annotate data points with events that could have explained fluctuations in self-reported levels of that construct. In addition, the constructs of mood, sleep, worry, enjoyment and energy might not have been relevant to all participants. Suggesting that other constructs should have been considered, or that participants should have been able to personalise their tracking. This was supported by the answers given to some of the questions in the post-study acceptance survey.

Another possibility is that participants in the PPI sessions were not representative of participants in the feasibility study. Particularly if we consider that pregnant women are a very heterogeneous group with just one factor in common: pregnancy. Suggesting that mental health needs during pregnancy can vary across a wide range of personal and socio-demographic characteristics, and that more effort is needed to ensure that, in practice, the technology is delivered to the intended target audience.

Low adherence rates could also reflect the de-contextualisation of BrightSelf. Although intended to be used as part of antenatal care, answers to the post-study acceptance survey suggest that this system was perceived as a research tool. Moreover, anecdotal evidence indicates that data collected through BrightSelf were not discussed during the antenatal appointments. In addition, other than the actions taken in the event of an alert, direct regular communication with midwives or any other member of the clinical team was not supported through BrightSelf. This de-contextualisation could have reduced participants' motivation to engage with this system.

Taken together, these findings highlight the need for healthcare providers to realign their views on patient adherence. Smartphones (and other forms of mobile technology) are intimate devices, used in real life and to support many aspects of an individual's life (one of which might or might not be their health). Developing mHealth interventions on the basis of traditional probability estimates of adherence negates these realities, thus leading to poor adoption [13]. To avoid this pitfall, healthcare providers might need to draw from other industries in their approach to technology development. Approaches that might be more compatible with the notion of patients as active entities in their healthcare.

In other industries, technology is developed iteratively, with multiple versions of a product being deployed over time. User-led co-design is an essential part of this process to ensure that technical specifications remain current and are aligned with the goals of users, and that user feedback is incorporated throughout. Moreover, minimally viable products that lead to more mature versions of a system are not uncommon. This is in contrast with the rather linear approach usually adopted in healthcare, where user-led co-design represents a step towards the development of 'finished products' [16, 17], professional expertise is prioritised over that of patients, and user requirements are assumed to be static and are often based on a relatively small sample of users.

Denying the iterative and user-centric nature of technology restricts the ability to develop interventions that are relevant, personalised and adaptive to changes in personal circumstances. Particularly when mobile technology could in theory gather information about the way patients interact with an intervention, enabling the implementation of technical or human support strategies to maintain and promote adherence (for example, assessing how the content, timing and frequency of push notifications affect their ability to promote engagement [18, 19]). It must be acknowledged however, that incorporating this approach to innovation in healthcare is an extremely complex process that requires addressing multiple system-wide barriers such as configuration of services, leadership, incentives schemes, commissioning and regulatory frameworks, and evidence generation amongst others [17, 20].

In addition, in realigning their notions of adherence, healthcare providers might need to address differences in user needs between patients and clinicians. Data that support the needs of patients might not be appropriate for clinical decision-making. As in the case of the

momentary data collected in this study. Perhaps mHealth interventions need to incorporate strategies that use data as a way of facilitating a meaningful dialogue between patients and clinicians, rather than as a way of obtaining clinical data summaries. In devising these strategies, it is crucial to ensure that the volume and types of data that patients bring to the consultation do not hinder clinicians' ability to make decisions. It might also be important to re-value non-adherence, not as a sign of lack of effectiveness, but as an indicator that patients are doing well and do not need the support of an mHealth system.

6.6.2. Implementation challenges

The proportion of participants who did not download, install and activate BrightSelf, and the alerts triggered during the study provided an opportunity to explore some of the implementation challenges associated with the deployment of mobile technology in clinical contexts.

The sites where more than half of participants did not download, install and activate BrightSelf highlighted issues around the lack of basic infrastructure. These sites had expressed concerns about not having access to a public wireless network. Instead of going through the installation process at the clinic, participants recruited in these sites were asked to download, install and activate BrightSelf at home. Thus reducing the likelihood of them completing this task. This indicates that healthcare providers wishing to 'prescribe' apps need to consider the basic infrastructure needed to deliver these directly onto patients' smartphones.

The alerts triggered during the study highlighted the need to monitor the potential impact that the introduction of a new technology can have on the workload of clinical teams. The proportion of alerts corresponded to the expected rates of depression in the antenatal period [21-24]. Therefore, the use of an app for the assessment of symptoms of depression in naturalistic settings might not result in a disproportionate increase in the number of pregnant showing an increased risk of depression. Although these results need to be interpreted with caution given the low levels of adherence with BrightSelf.

Furthermore, the types of alert provided further insights into the impact that BrightSelf had on workload. Although orange alerts represented the most common type of alert, they were

not a source of major concern due to the proportion of participants that triggered alerts. The higher proportion of yellow alerts in the *Retrospective plus momentary assessments* group was explained in terms of the increased sampling frequency and the not so clear distinction between the *Check Back* and the *Check In* features. This could have led to the completion of more than one daily EPDS assessment, which was the main criterion for triggering this type of alert. Moreover, the higher proportion of red alerts in the *Retrospective assessments* group might suggest better patient engagement with the data collection. All these possibilities need to be considered when devising data processing protocols than complement the incorporation of a new technology into a clinic's workflows.

6.6.3. Limitations

The main limitations of this study was the de-contextualisation of BrightSelf. In order to fully evaluate the role of consumer mobile devices for the assessment and monitoring of mood and depression during pregnancy it is crucial to ensure that these technologies are embedded into existing workflows and care pathways.

Moreover, the main features of BrightSelf were the *Check In* and the *Check back* features. Although recommendations for low intensity activities were provided through the app, these were not linked to the data that were inputted. Instead, it relied on participants using that feature voluntarily. In addition, the completion of assessments was not linked to personalised advise or recommendations. Although these decisions were based on the concerns raised by midwives during the PPI sessions, they might have contributed to poor engagement with, and the de-contextualisation of, the app.

As with the feasibility study presented in Chapter 3, the study procedures were administered by research staff with no links to the clinical care team or by midwives with dual responsibilities. Moreover, the administrative research requirements were unnatural to the realities of antenatal care. The combination of these factors might have contributed to the de-contextualisation of BrightSelf.

Another important limitation of this study was the inability to explore at a very granular level the reasons for missing data. This could have provided useful information about the reasons

why adherence in a particular group of participants was low, leading to refinements in BrightSelf.

6.6.4. Conclusions

This study suggests that it is feasible and acceptable for pregnant women to use bespoke mobile apps running on their own devices for the longitudinal assessment and monitoring of mood and depression during pregnancy in the context of antenatal care provided through the NHS. The incorporation of momentary assessments does not appear to improve participant adherence. This however, could be the result of the inability of BrightSelf to meet the needs of its intended audience; heterogeneity in the target population in terms of their individual characteristics and needs; or the de-contextualisation of BrightSelf.

In order to maximise long-term adherence with such systems, it is important to make design choices (both in terms of software development and sampling protocols) that are sensitive to the needs and goals of users (whether it is clinicians, patients or both). It must be acknowledged however, that these needs and goals might vary considerable within any given target population. In addition, strategies to promote adherence must reflect individual needs, and should be adaptive to changes in personal circumstances. This requires an iterative approach to the design, development and deployment of mHealth systems.

Moreover, these systems provide healthcare providers and researchers with an opportunity to evaluate patterns of engagement, thus facilitating the segmentation of target populations and the personalisation of strategies to promote engagement. Consumer mobile devices also offer the opportunity to adopt a more proactive approach to the provision of antenatal mental health care. This however, would require careful consideration of how services are configured.

Overall, realising these opportunities require a realignment of healthcare providers' views on patient adherence, and on the adoption of an iterative and user-centric approach to technology development in healthcare.

References (Chapter 6)

1. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, Merchant GC, Naughton F, Blandford A. Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *Americal Journal of Preventive Medicine.* 2016;51(5):833-842. Available from: doi: 10.1016/j.amepre.2016.06.015 [Accessed 23 July 2018].
2. Marcano-Belisario JS, Jamsek J, Huckvale K, O'Donoghue J, Morrison CP, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. *Cochrane Database of Systematic Reviews*. 2015;7:MR000042. Available from: doi: 10.1002/14651858.MR000042.pub2 [Accessed 24 September 2018].
3. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence-based mental health*. 2018;21:116 – 119. Available from: doi: 10.1136/eb-2018-102891 [Accessed 24 September 2018].
4. Doherty K, Barry M, Marcano-Belisario JS, Arnaud B, Morrison C, Car J, Doherty G. A Mobile App for the Self-Report of Psychological Well-Being During Pregnancy (BrightSelf): Qualitative Design Study. *JMIR Mental Health*. 2018;5(4): e10007. Available from: doi: 10.2196/10007 [Accessed 01 January 2019].
5. Ahmed R, Aslani P. What is patient adherence? A terminology overview. *International Journal of Clinical Pharmacy*. 2014;36(1): 4 – 7. Available from: doi: 10.1007/s11096-013-9856-y [Accessed 26 September 2018].
6. Morrison C, Doherty G. Analyzing Engagement in a Web-Based Intervention Platform Through Visualizing Log-Data. *Journal of Medical Internet Research.* 2014;16(11):e252. Available from doi: 10.2196/jmir.3575 [Accessed 02 October 2018].
7. Marcano-Belisario JS, Doherty K, O'Donoghue J, Ramchandani P, Majeed A, Doherty G, Morrison C, Car J. A bespoke mobile application for the longitudinal assessment of depression and mood during pregnancy: protocol of a feasibility study. *BMJ Open.* 2017;7: e014469. Available from: doi: 10.1136/bmjopen-2016-014469 [Accessed 24 June 2018].

8. Schwartz JE, Stone AA. Strategies for analyzing ecological momentary assessment data. *Health Psychology*. 1998;17(1):6-16.
9. Shiffman S. Conceptualizing analyses of ecological momentary assessment data. *Nicotine & Tobacco Research*. 2014;16(Suppl 2):S76-87. Available from: doi: 10.1093/ntr/ntt195 [Accessed 12 May 2017].
10. Hesser H. Modelling individual differences in randomized experiments using growth models: Recommendations for design, statistical analysis and reporting of results of internet interventions. *Internet Interventions*. 2015;2(2):110-120. Available from: doi: 10.1016/j.invent.2015.02.003 [Accessed 20 July 2018].
11. Garcia TP, Marder K. Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Diseases as a Model. *Current Neurology and Neuroscience Reports*. 2017;17(2):14. Available from: doi:10.1007/s11910-017-0723-4 [Accessed 20 July 2018].
12. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. Available from: doi: 10.1136/bmj.b2393 [Accessed 20 July 2018].
13. Neff G, Nafus D. *Self-Tracking*. London, UK: MIT Press; 2016.
14. Torous J, Nicholas J, Larsen ME, Firth J, Christensen. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence Based Mental Health*. 2018;21(3):116-119. Available from: doi: 10.1136/eb-2018-102891 [Accessed 04 October 2018].
15. Torous J, Chan SR, Tan SYM, Behrens J, Mathew I, Conrad EJ, Hinton L, Yellowlees P, Keshavan M. Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics. *JMIR Mental Health*. 2014;1(1): e5. Available from: doi: 10.2196/mental.4004 [Accessed 12 May 2017].
16. Castle-Clarke S, Edwards N, Buckingham H (2017). *Falling short: Why the NHS is still struggling to make the most of new innovations*. [Online] Nuffield Trust – Evidence for better health care. Available from: https://www.nuffieldtrust.org.uk/research/falling-short-why-the-nhs-is-still-struggling-to-make-the-most-of-new-innovations [Accessed 01 January 2019].

17. Imison C, Castle-Clarke S, Watson R, Edwards N (2016). *Delivering the benefits of digital health care*. [Online] Nuffield Trust – Evidence for better health care. Available from: https://www.nuffieldtrust.org.uk/research/delivering-the-benefits-of-digital-health-care [Accessed 01 January 2019].

18. Bidargaddi N, Pituch T, Maaieh H, Short C, Strecher V. Predicting which type of push notification content motivates users to engage in a self-monitoring app. *Preventive Medicine Reports*. 2018; 11: 267 – 273. Available from: doi: 10.1016/j.pmedr.2018.07.004 [Accessed 4 October 2018].

19. Morrison LG, Hargood C, Pejovic V, Geraghty AWA, Lloyd S, Goodman N, Michaelides DT, Weston A, Musolesi M, Weal MJ, Yardley L. The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone-Based Stress Management Intervention: An Exploratory Trial. *PLOS One*. 2017; 12(1): e0169162. Available from: doi: 10.1371/journal.pone.0169162 [Accessed 4 October 2018].

20. Accelerated Access: Review of innovative medicines and medical technologies supported by Wellcome Trust (2016). *Accelerated Access Review: Final Report*. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/565072/AAR_final.pdf [Accessed 01 January 2019].

21. Gavin NI, Gaynes BN, Lohr KN, Meltzer-Brody S, Gartlehner G, Swinson T. Perinatal depression: a systematic review of prevalence and incidence. *Obstetrics and Gynecology*. 2005; 106(5 Pt 1): 1071-83.

22. Dagklis T, Papazisis G, Tsakiridis I, Chouliara F, Mamopoulos A, Rousso D. Prevalence of antenatal depression and associated factors among pregnant women hospitalized in a high-risk pregnancy unit in Greece. *Social Psychiatry Psychiatric Epidemiology*. 2016;51(7): 1025-31. Available from: doi: 10.1007/s00127-016-1230-7 [Accessed 4 October 2018].

23. Bennett HA, Einarson A, Taddio A, Koren G, Einarson TG. Prevalence of depression during pregnancy: systematic review. *Obstetrics and Gynecology*. 2004;103(4): 698-709.

24. Parker GB, Hegarty B, Paterson A, Hadzi-Pavlovic D, Granville-Smith I, Gokiert A. Predictors of post-natal depression are shaped distinctly by the measure of

'depression'. *Journal of Affective Disorders*. 2015;173: 239-44. Available from: doi: 10.1016/j.jad.2014.10.066 [Accessed 4 October 2018].

7. Discussion

7.1. Overview

This doctoral project evaluated the effect of mobile apps on the quality of data collected through antenatal depression screening scales (i.e., delivery mode effects). Delivery mode effects were considered in relation to the influence of apps on survey responses and on the survey completion process. Moreover, delivery mode effects were compared across two data collection settings: clinical and naturalistic.

The findings of this project show that the setting in which surveys are completed will determine what data quality domains are relevant and how they should be defined, as well as the contextual cues that influence patients' responses. The findings also demonstrated that the adaptation of validated surveys for use with apps should include an evaluation of factors, such as survey layout, that can mediate survey responses. Both the setting of survey completion and the factors that influence survey responses have theoretical and methodological implications for the validation of surveys, as well as for how patient adherence with digital health technologies is understood. Moreover, the findings of this project have wider implications for the design of technologies that must reconcile the often-competing needs of patients and practitioners, be compatible with the clinical contexts in which they are implemented and embed themselves within the routines of their end users.

Overall, mobile technology should be viewed as powerful tool for supporting and complementing the role of healthcare practitioners, empowering patients, and enabling the reconfiguration of health services to better meet the needs of their users.

7.2. Rationale

Mobile technology is considered one of the key technological tools for improving many aspects of healthcare, including the collection of patient data. This perception is often based on idealistic notions about the power of technology to solve complex problems but fail to consider individual contexts of patient interaction with technology and the wider clinical contexts within which technology is meant to operate. The aim of this doctoral research was to evaluate how these contexts influence the completion of health-related surveys via apps, and thus the quality of the corresponding responses. Survey responses were the main focus of this work given their role in informing many diagnostic and therapeutic decisions.

7.3. Discussion

This doctoral work evaluated the impact that mobile apps had on the quality of self-administered survey responses through a systematic review (Chapter 2) and two feasibility studies about the use of apps for the assessment of mood and symptoms of depressions in clinical (Chapter 3 and 4) and non-clinical settings (Chapters 5 and 6) within the context of antenatal care. This chapter outlines the key contributions of this work and discusses its wider implications for clinical practice and research. For an overview of findings see Appendix 38.

7.3.1. Key contributions

The main contributions of this doctoral work were:

- recognising the influence that the setting of data collection has on data quality;
- recognising the diversity of factors that can lead to, and mediate, delivery mode effects;
- acknowledging that the setting of data collection and the mediators of delivery mode effects can influence how the validation of surveys is understood; and
- understanding how the setting of data collection can influence the notions of patient adherence with digital health technologies.

These contributions are discussed in the following sub-sections and summarised through a conceptual framework presented at the end of this section.

7.3.1.1. Setting of data collection and its influence of data quality

Setting of data collection refers to the location(s) and situational context(s) in which the completion of self-administered surveys occurs [1]. Although relevant to mobile apps, the findings of this doctoral work have the potential to be extended to other forms of technology-enabled data collection such as wearable sensors, for example. A key contribution of this thesis was to identify the main data collection settings and to determine how they influence data quality.

The systematic review discussed in Chapter 2 identified two main data collection settings: clinical and naturalistic [1-4]. Each of these is best considered as an umbrella term for a wide range of locations and situational contexts. As discussed in Chapters 2 and 3, clinical settings refer to data collection that occurs in clinical facilities [1-3]. Encounters in these facilities are part of an established patient-healthcare provider dynamic, with clear roles and expectations. These facilities however, vary in terms of level of care (i.e., primary, secondary, tertiary and community), clinical focus (e.g., antenatal care, mental health, or general practice), type of consultation (e.g., outpatient vs inpatient care), and the healthcare professional leading the consultation (e.g., midwife, imaging technician, or GP). Therefore, although the patient-healthcare provider dynamic provides a sense of homogeneity, the variation discussed here can cause situational contexts to be heterogeneous.

Naturalistic settings refer to data collection that occurs outside clinical facilities [1, 4]. Although broad in its scope, this category is important as it represents patients' lived experience. These settings include a wide range of locations that vary in terms of the presence and stability of pre-established dynamics and the situational contexts they provide. For example, work usually provides a stable employer-employee dynamic with clear roles and responsibilities, which is arguably similar to the patient-healthcare provider dynamic. Moreover, some locations might enable temporal variation in their situational contexts. Therefore, the influence of naturalistic settings on data collection can be much more heterogeneous than clinical settings.

Which setting is chosen depends on clinical need. For example, if the aim is to inform a clinical decision during a consultation on the basis of a single administration of a validated scale, then

a clinical setting will be more appropriate. If, on the other hand, the aim is to assess treatment response over time, then the repeated administration of surveys in naturalistic settings will be more appropriate. The choice of setting can have a direct and an indirect influence on data quality. Its direct influence is related to the characteristics of the data collection process (e.g., its frequency and duration), which determine what data quality domains are relevant and how they should be defined.

Its indirect influence is related to the situational contexts that are in operation during data collection, which determine the cues that influence patients' responses [5, 6]. This is the reason why the heterogeneity of the setting of data collection is relevant, as different cues might get activated throughout the data collection process. These cues however, remain latent until the technology triggers an instance of data collection. Data collection settings also exert their indirect influence on data quality through the technology itself; however, this will be discussed in the next sub-section.

An example might help to illustrate this discussion. In clinical settings, data completeness can be defined in terms of survey items (e.g., making sure that patients answer all the EPDS questions) [1-3]. In this setting, an app might need to run on devices owned by healthcare providers, offer automatic scoring of surveys and support offline data collection. In addition, validation rules might be sufficient to ensure data completeness and patients might feel more comfortable disclosing sensitive information on a tablet computer. In addition to survey items, data completeness in naturalistic settings can also be defined as a temporal construct (e.g., making sure that participants complete monthly assessments for six months) [4]. In these settings the technical requirements will be different: an app might need to run on patients' own devices and enable automatic data transfer, and validation rules might need to be complemented with notifications and reminders. Moreover, completing a health survey might be acceptable at home but not whilst commuting or when at the theatre. This example shows how the same outcome and the same technology might be operationalised differently according to the setting of data collection, which leads to different quality assurance strategies.

7.3.1.2. Mediators of delivery mode effects

Delivery mode effects refer to changes in the properties of survey responses and other data quality domains that are associated with the mode through which a survey is delivered [1, 5, 6, 7-12]. As discussed in Chapter 2, the size and direction of these effects can be mediated by the properties of the technology, or by the way in which the technology interacts with respondents or with the setting of data collection. A key contribution of this thesis was to assess delivery mode effects and its mediators in relation to validated clinical surveys.

The use of apps as a delivery mode for clinical surveys is often based on overall assessments of data equivalence, as shown by the systematic review discussed in Chapter 2 [1]. Moreover, guidance in this area indicates that as long as no significant changes are made to the content of a survey, its properties should remain unchanged regardless of the delivery mode [13]. However, findings from psychology and survey methodology have shown that decisions about survey design, human computer interaction factors and choice of technology affect the responses given to general purpose surveys [7-12]. Replicating these effects in healthcare is important given the potential for bias in clinical decisions.

Only survey layout and assessment strategy were assessed in this thesis as mediators of delivery mode effects in clinical and naturalistic settings, respectively [2-4]. Although survey layout affected the proportion of participants scoring above a recommended clinical threshold [3], the effect of an assessment strategy consisting of momentary questions could not be assessed due to high levels of missing data (although there is sufficient theoretical justification to continue exploring the effect of momentary data on retrospective assessments [14-18]). What this work demonstrates is that deploying new technology for the administration of validated clinical scales on the basis of overall equivalence studies could lead to unintended delivery mode effects. This highlights the need for an exhaustive theory-driven map of the potential mediators of delivery mode effects, and the evaluation of their impact on clinical scales in multiple settings.

7.3.1.3. Validation of surveys

Validation is a crucial part of the quantitative survey method [5, 6, 19]. Establishing the validity, reliability and other diagnostic properties of a survey helps to ensure that data are accurate, appropriate for the intended purpose, and that can be standardised across patients and sites [5, 6]. Validation is based on the performance of survey items under specific circumstances and in particular settings. It also requires the aggregation of results from multiple studies [5]. A key contribution of this thesis was to show how apps provide new practical and methodological opportunities for validation methods.

From a practical perspective, apps can facilitate the data management requirements of validation studies. Indeed, Chapter 4 showed how apps can maximise the scalability and speed of data collection and analysis for assessing the diagnostic properties of self-administered clinical surveys. In addition, apps could minimise the temporal and geographical constrains on the administration of surveys in naturalistic settings. They could also expand the number of personal and contextual variables that can be captured.

These practical advantages present new challenges and new opportunities for validation methods. Expanding the settings of data collection multiplies the number of factors that can influence responses, as well as the sources of heterogeneity [20]. Although designing surveys that can reliably measure the constructs of interest under these circumstances becomes a challenge, the data collection capabilities of apps can help to better contextualise responses. This could be achieved through the collection of momentary data, for example [4]. Therefore, apps could increase knowledge about the dimensions associated with health constructs (i.e., content validity), reveal additional variables that are correlated with these constructs (i.e., criterion and discriminant validity), and help to identify response patterns across locations and time.

Coupled with multiple patterns of utilisation, apps could facilitate the discovery of new phenotypes with better predictive power, rather than continue to confirm already established phenotypes [21, 22]. Likewise, apps could help to identify patients' unique baseline patterns and to establish personal threshold values [22]. Overall, apps and other

forms of digital health can enable an approach to validation that focuses on individual rather than on generic probabilities [22].

7.3.1.4. Notions of patient adherence

Engagement with digital health interventions has been identified as one of the key factors for their effectiveness, and is usually defined as patient adherence [23]. As discussed in Chapter 5, adherence is commonly understood in relation to patients as passive recipients of expert wisdom who ought to comply with the requests of healthcare practitioners [24]. This is contrary to the views of users as active and independent entities that are common in the technology sector. A key contribution of this thesis was to highlight that medical notions of adherence need to be aligned with technological notions of user engagement if digital health is to be more impactful.

Adherence in clinical settings does not represent a real problem, as data collection typically consists of the single administration of a survey and survey completion is a common component of the situational context [1]. Moreover, the role of healthcare providers in these settings, and hence their needs, is more prominent. They also have more control over strategies to promote compliance with requests for data. Therefore, it is reasonable to expect good levels of patient adherence with digital health tools in clinical settings. However, the same cannot be expected in naturalistic settings.

Adherence in naturalistic settings is more relevant given that data collection is usually longitudinal. In these settings however, adherence has traditionally been very low [23, 25]. As discussed in Chapter 5, to promote adherence, digital health interventions need to be compatible with the individual contexts of use and need to be able to support patients' needs (as well as the needs of clinicians if the intervention is embedded in a clinical context) [23, 25, 26]. This is not easy to achieve due to the inherent heterogeneity of personal and contextual circumstances.

For these reasons, in naturalistic settings, it is more realistic to expect clusters of different patterns of adherence. This is compatible with current notions of adherence in which patients are considered active entities in their own healthcare. As such, this conceptualisation of

adherence is similar to the view commonly held in the technology sector of users as active and independent entities. The importance of realigning traditional views of adherence becomes obvious when one considers the current policy debate that is increasingly shifting the locus of responsibility towards patients, and that promotes the use of digital technology as the means to achieve positive change in healthcare [27-29].

Therefore, the development and implementation of digital health interventions in naturalistic settings ought to consider the personal circumstances of patients, the locations and situational contexts in which the technology will be used, the goals that the intervention is supposed to support, and the wider clinical context. More importantly, it ought to consider individual patterns of actual use. One way of achieving this is to adopt an iterative and user-centric approach to technology development and evidence generation.

These key contributions of this doctoral work are summarised in the conceptual framework shown in Figure 22. At the core of this diagram, and a common theme across the thesis, is data quality. A detailed argument about the importance of data quality in healthcare was provided in Chapter 1; in short, it was chosen as the main focus of this thesis due to its role in clinical decisions. To better understand how data quality can be affected when using digital health tools, it is helpful to consider the layers that represent the wider context of implementation.

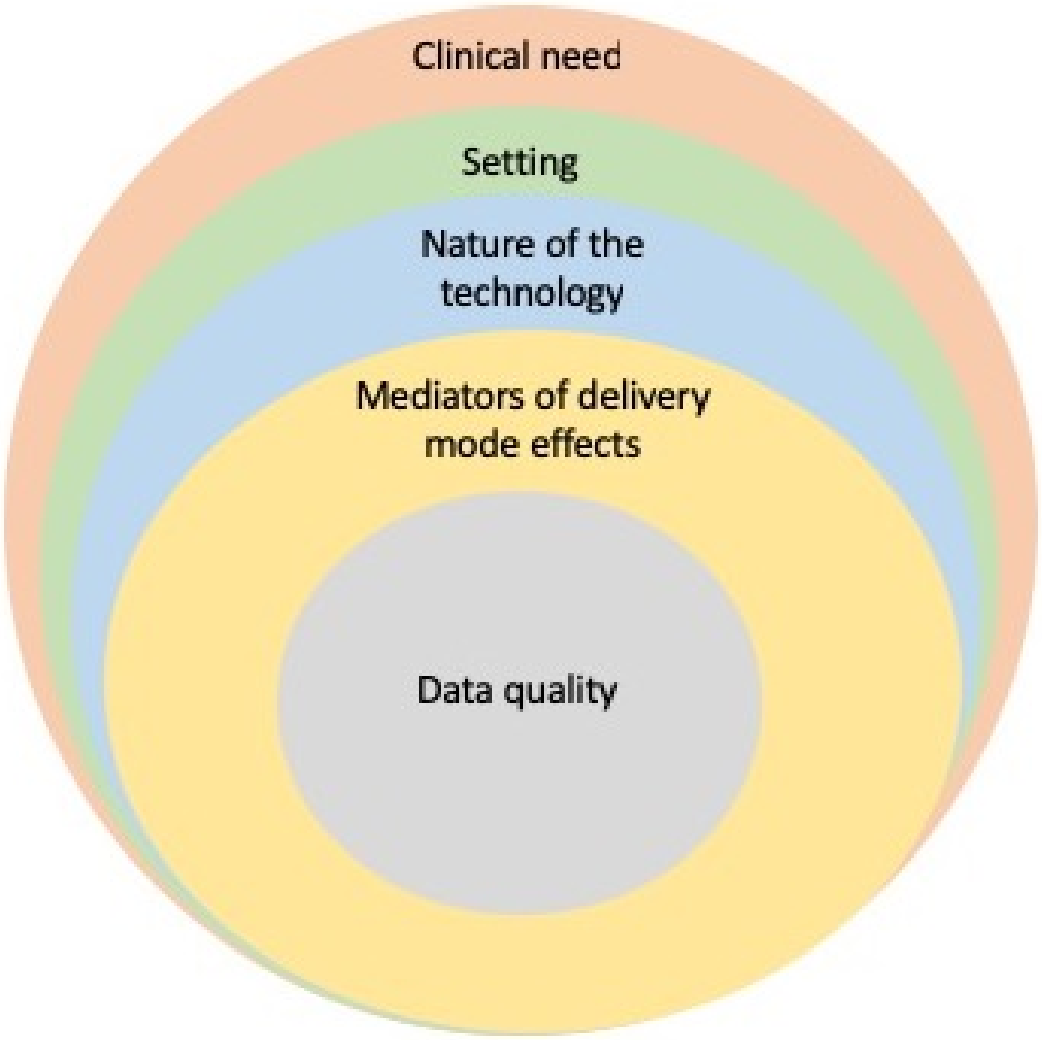


Figure 22. Conceptual framework of the effect that technology for the collection of clinical data has on data quality

Starting with the outer layer, clinical need refers to the problem that needs addressing. It involves the wider clinical context, key individuals, their goals, and the key actions needed to achieve them (including duration and frequency). The next layer refers to the setting in which key actions will be performed, and it includes both location and situational contexts. This then leads to the next layer, which refers to identifying the technology that is best placed to help individuals to perform their actions and achieve their goals. Further down the diagram, it becomes possible to identify those factors that could lead to, and mediate, delivery mode effects: survey design, human computer interaction and technical specifications of the device. Delivery mode effect could also be the result of the way in which the delivery mode interacts with the wider clinical context and with the setting in which the technology is being used. The

final layer refers to the data quality domains that are of interest, and how they should be defined and assessed.

An example can explain this diagram best. If one considers the study in Chapter 3, the main *clinical need* was to facilitate antenatal depression screening in antenatal clinics. Therefore, the key individuals were midwives wishing to assess the risk of depression in their patients, so that appropriate action could be taken during the consultation. For reasons outlined in previous chapters, the supporting action was to ask pregnant women to complete the EPDS in the waiting area prior to their appointment. This was considered in the wider context of antenatal care, which provides structured encounters between pregnant women and antenatal services, and clearly defined roles for both of them. Asking pregnant women to complete this scale fitted with the expectations of this encounter in a clinical *setting*. The *technology* chosen was a commercial app running on tablet computers, which could satisfy the needs of midwives and integrate itself into a clinic's workflows: automatically score women's answers to the EPDS, and provide a summary of their scores that could be communicated to the consulting midwife. This then determined the *mediators of delivery mode effects* such as survey layout, compatibility of survey completion with clinics' waiting times and acceptability of the technology, for example. As a result of this, the relevant *data quality dimensions* that were assessed included differences in EPDS responses, completion time, and breakoff rates.

One advantages of this conceptual framework is that it helps to identify the points at which data quality is likely to be affected. As a result of this, it might be possible to devise strategies that can minimise a potential detrimental effect. These strategies might involve technology (e.g., programmable reminders or data-driven personalised actions), or they might involve the wider clinical context (e.g., offering additional human support through the telephone when needed, or ensuring that the data collected are discussed during a consultation).

Another advantage of this conceptual framework is that it helps to think about the clinical problem and the appropriateness of a technology-based solution. This is important to avoid the pitfall of thinking that technology alone can solve complex problems, or that technology is always best [22]. The fact that data on a health-related construct can be collected does not necessarily mean that this will translate into clinical benefits for patients. Perhaps these data

(or the technology itself) might not be acceptable or helpful to patients or clinicians. Moreover, if a technology-based solution is deemed appropriate, this framework helps to contextualise it. This is important to ensure a good fit between the technology, and the personal circumstances of patients and/or the workflows of clinical care teams.

7.4. Implications for clinical practice

7.4.1. Workload and workflows: the importance of external validity

The findings from this doctoral work indicate antenatal services could rely on the use of mobile apps for administering self-reported health surveys. In the case of depression screening, this delivery mode might not lead to a disproportionate number of pregnant women screening positive as it is sometimes feared. The privacy afforded by apps is sometimes perceived as encouraging users to disclose information that they would otherwise not feel comfortable discussing with another person. In this work however, the proportion of pregnant women who scored above the established EPDS thresholds was always consistent with the expected prevalence of antenatal depression.

Nonetheless, healthcare providers should always be mindful of the potential impact that apps can have on their workload and workflows. Perhaps the roll out of mobile technology in clinics should be preceded by pilot studies that can help to identify these issues. These pilot studies need to be different from traditional academic studies that prioritise internal validity over external validity. Instead, pilot studies should reflect the local idiosyncrasies of a clinic or clinical service. Consequently, healthcare providers should understand that the findings from pilot studies in other settings might not be applicable to their own setting. Therefore, they need to identify the lessons that are transferable due to similarities in service configuration and modify those aspects of a pilot study that are not relevant to them.

7.4.2. Adaptation of a survey for use with an app: data quality

Healthcare providers must also acknowledge the complexities that adapting a health survey for use with an app entail. Particularly, in relation to any potential impact on the quality of the data gathered. This work showed that a seemingly simple modification in the EPDS layout can result in a delivery mode effect [2, 3]. Indeed, this could be the case for other adaptations.

This might not be too problematic when survey completion occurs in a clinic and staff members can arguably react to the data promptly. However, delivery mode effects could become more serious when patients are expected to use mobile apps on their own in naturalistic settings. Therefore, when adapting surveys for use with an app, healthcare providers should rigorously document all the changes made to such a survey. They should also keep a strict version control of all the surveys, and monitor patient responses. It might also be beneficial to work alongside researchers who have experience evaluating delivery mode effects.

7.4.3. Integration of patient data into a consultation

Healthcare providers wishing to deploy mobile technology in their clinics need to consider the integration capabilities offered by software vendors. Especially since they are unlikely to have the resources and the expertise necessary to develop and implement their own bespoke systems. This does not imply that healthcare providers could not implement other strategies for using data effectively in the absence of inter-operability. In the study discussed in Chapter 3, this took the form of a letter that summarised a patient's *results* on the Whooley Questions and the EPDS. Although this approach might seem to defeat the point of using mobile technology, it actually enabled the administration of screening scales ahead of the consultation and the automation of the scoring procedures. In theory, this enabled clinics to implement NICE recommendations for the recognition of antenatal depression without having to use consultation time for data gathering and processing. This assumption needs to be explored further as this work did not include a direct assessment of how the letters were used during the consultations. Nonetheless, it highlights that healthcare providers could rethink their workflows in a way that allows them to gather data before the consultation and use the consultation time to have a more meaningful dialogue with a patient.

7.4.4. Data processing and data visualisation

When using apps for gathering patient data, healthcare providers need to consider how the data will be processed and visualised. This is to their benefit, as well as to the benefit of their patients. Practitioners might use these data to inform diagnostic or therapeutic decisions. However, it is unrealistic to expect them to review individual data points (e.g., each question

in the EPDS). Instead, the computing capabilities of consumer mobile devices can be harnessed to meet the scoring requirements of validated scales, or any other data processing requirements. In this way, practitioners can be presented with data summaries that contain relevant and actionable information, in a format that they understand [30].

The requirements for patients might be slightly different, particularly if they are expected to use an app on their own over a prolonged period of time. Although useful for healthcare professional, clinical scales such as the EPDS might not be appropriate for the needs of patients. Therefore, other self-reported measures might be required. However, this can present certain challenges. Not only measures need to resonate with the experience of patients, the data summaries must be meaningful and allow them to take action when necessary [30]. The implication is that patients need to engage in a dialogue with their healthcare providers before deciding on what domains to measure and how. There is then a question of how to create apps with this level of built-in flexibility, and around the responsibility for the creation of such apps. Healthcare providers might not have the resources or the expertise necessary to achieve this. Nor is the current commissioning landscape conducive to ongoing agreements between NHS organisations and software developers that reflect the evolving nature of software. Instead, commissioning of digital products in the NHS is treated similar to other services and products: static products that meet a pre-specified list of requirements. This highlights the need to explore new ways in which digital health services (including mobile apps) must be commissioned.

Alternatively, patients can choose from the myriad of apps available in the current marketplaces. This raises questions around the lack of standardisation in how data are collected, processed and visualised. Practitioners with patients bringing third-party apps into the consultation would need to deal with multiple forms of data, some of which might not be relevant. Not to mention the implications for patient safety and quality of care when using unregulated products. Therefore, policy makers might need to work towards the creation of standards (e.g., data dictionaries, security certification) that app developers should follow when creating a health app.

7.4.5. Locus of responsibility

Apps (and other digital health interventions) can facilitate and support self-monitoring and self-management activities, which could lead to patient empowerment and better clinical outcomes. However, they could have the opposite effect and result in patients having to bear the full responsibility of their own healthcare. For example, patients might be expected to take action on the basis of unhelpful or uninformative data summaries provided by an app. This could be extremely problematic if patients do not have the knowledge and the skills needed to perform complex self-management tasks. For this reason, healthcare providers and policy makers need to ensure that digital health is used as a tool to complement and enhance healthcare delivery, and not to replace it as part of a cost-cutting drive.

7.4.6. Provision of basic infrastructure and information governance

If healthcare providers (and policy makers to an extent) expect apps to become an important part of the medical arsenal, they need to provide the basic infrastructure to deliver these apps. When data are collected in clinical settings, they might need to consider providing devices onto which patients can complete surveys. This presents certain challenges associated with the physical security of these devices (e.g., theft), and with the security of the data held in them. These issues must be reflected in relevant information governance policies.

If patients are expected to download and install an app onto their own handsets, then clinics should provide access to a Wi-Fi network. In this way, patients will not incur the financial costs associated with data usage. In addition, it will help practitioners to ensure that apps are actually downloaded and installed, and to set the context in which they will be used. It also creates an opportunity for patients to ask questions or raise concerns about using an app.

7.4.7. Diversity and inequalities

Healthcare systems are typically configured around individual diseases or around groups of diseases affecting a physiological or anatomical system. This introduces heterogeneity in terms of patient populations that differ in their pathophysiology, therapeutic options, self-management requirements, need and goals. However, most diagnostic labels are applied to a common disease phenotype that could be associated with a multitude of aetiological

factors. Or they might just reflect the lack of measurable indicators that could help to differentiate these seemingly identical phenotypes. Regardless, the implication is that there is considerable heterogeneity even within a single patient population.

For example, this doctoral work focused on women with uncomplicated pregnancies with no current medical history of depression or anxiety, who were attending antenatal care. In terms of mental health, it is possible that women with complicated pregnancies might have a different baseline risk for depression, as well as different psychological well-being needs. These factors might even vary depending on the type of complication. Similarly, women with current history of depression will also have different needs as they are already dealing with this condition. Moreover, the sample of participants was predominantly white, educated to degree level and in employment. These and other socio-demographic factors can influence a woman's baseline risk for antenatal depression. Not to mention other risk factors such as the level of social and family support, experiences of domestic abuse, substance abuse, or the co-existence of other mental health conditions.

This example helps to illustrate an important point about digital health. Digital health has been heralded as one of the key components needed to achieve personalised healthcare. To realise this goal, healthcare practitioners and policy makers (amongst other stakeholders) must embrace the intrinsic heterogeneity that was just described. This requires rethinking the way healthcare systems are configured and the way digital health systems are commissioned. What might work for one patient (or group of patients) might not apply to others. In addition, the ability to gather more data (both in terms of number of data points, as well as number of variables) can lead to entirely new disease phenotypes. This is before we acknowledge that patient circumstances can also change over time. The risk is the implementation of digital health solutions on the basis of evidence obtained from a non-representative group of individuals. These solutions might then fail to meet the needs of most patients. If these patients happen to be those at most risk, then digital health can contribute to widen existing health inequalities.

7.5. Implications for research

7.5.1. Delivery mode effects

The adaptation of a survey for use with a mobile app is a complex process that involves decisions about the setting in which surveys will be completed, survey design, presentation of questions, data entry formats, and human computer interaction factors amongst others [1, 7-12]. These decisions have the potential to affect the quality of survey responses either directly or indirectly through changes to the survey completion process. This phenomenon is known as delivery mode effect [5, 6], and it can have important implications for clinical decisions that are based on survey responses.

The healthcare literature is mainly focused on studies that consider the overall data equivalence between survey responses obtained via an app, and those obtained via the original delivery mode. In addition, these studies do not consider the possibility of delivery mode effects as a result of decisions made during the adaptation process. Similar to the effect of survey layout shown in Chapter 3 [3], other factors could lead to delivery mode effects when using validated health surveys. Therefore, researchers need to better understand how specific adaptations affect the overall properties of a survey, as well as item responses.

7.5.2. Importance of survey completion setting

Apps could be an effective delivery mode for self-administered surveys in both clinical and naturalistic settings [1]. However, these settings differ in important ways. For example, survey completion in clinical settings typically consists in the single administration of one or more usually validated surveys. In addition, survey responses tend to be integrated within a consultation almost immediately. Survey completion in naturalistic settings typically consists in the repeated administration of one or more surveys over an extended period of time. Here, non-validated surveys are more common, and responses need to be aggregated and summarised before they can be integrated within a consultation.

The choice of endpoints when planning research studies must reflect these differences between survey completion settings. For example, overall data equivalence, completion time, and breakoff rates are relevant in clinical settings. They provide a good indication of the

impact that apps can have on the properties of a validated survey and on the workflows of a clinic. However, in naturalistic settings these indicators become less relevant. Instead, patient adherence and data completeness become more important.

Moreover, naturalistic settings comprise a wide range of locations and situational contexts that differ in terms of the factors that could influence survey responses. This can make validation attempts challenging given the differences between and within participants in the circumstances under which each instance of a survey is completed. This requires new approaches to the validation of self-administered scales.

7.5.3. Evidence-based implementation

Clinics are complex entities comprising multiple stakeholders, workflows and pathways. In addition, there are regional differences in how clinics operate. The deployment of mobile technology in these clinics require the implementation of strategies that are sensitive to these local idiosyncrasies. Researchers should consider methodologies such as *implementation science* when evaluating digital health interventions. Implementation science enables researchers to evaluate the effectiveness of strategies for the implementation of complex interventions with demonstrable efficacy and/or effectiveness. In the absence of efficacy or effectiveness, this methodology offers a series of hybrid designs that enable the evaluation of both implementation strategies and intervention components. This type of evidence might be more beneficial for practitioners wishing to adopt mHealth solutions.

7.5.4. Adherence

Patient adherence with a mHealth intervention is better conceptualised as a cluster of individualised patterns, rather than as a summative phenomenon [31, 32]. Therefore, researchers in this area should aim to identify these individualised patterns, and the personal and contextual factors associated with them. An obvious advantage of this approach is the development of interventions that are tailored to the individual, as envisioned by the proponents of personalised medicine. Although challenging in principle, this could be facilitated by the analysis of the log data and the built-in sensors data that consumer mobile devices are able to collect, or through the use of other wearables sensors.

Furthermore, human facilitation has been identified as one of the key factors that can improve patient adherence with digital health interventions [23]. Indeed, one of the proposed explanations for the low adherence rates observed in Chapter 6 was that the app was not used in the context of the interaction between pregnant women and their midwives. Therefore, researchers need to understand the strategies that could facilitate the incorporation of human facilitation in mHealth interventions, or the incorporation of mHealth interventions within the wider clinical context.

In addition, researchers need to understand what motivates patients to use a mHealth solution. User-led co-design techniques can help researchers to incorporate the views and experiences of patients when developing interventions and implementation strategies. However, researchers need to acknowledge that this is an iterative process, rather than the linear process common to healthcare research. Moreover, researchers should aim to capture as many diverse views as possible.

7.5.5. Pace of technology change

There is an obvious discrepancy between the pace of technology change and the lifecycle of academic research [33]. Particularly since RCTs are considered the gold standard in evidence-based medicine, and these studies are notoriously expensive and time consuming. In order to remain current, researchers in this area need to consider study designs that can accommodate the constant changes in technology development (both in terms of software and hardware development), as well as the sources of heterogeneity outlined previously. Methods such as implementation science, multiphase optimisation strategy, sequential multiple assignment RCT or realist methodology are ideal for this purpose. For this to happen however, incentives need to change. For example, funding bodies, research sponsors, commissioners and decision makers in healthcare organisations must widen the types of evidence that they find acceptable.

7.6. Limitations

For practical reasons, this project focused on mobile apps; however, the potential of mHealth extends further. Smartphones and tablet computers can log usage data and are equipped

with sensors that can capture users' behaviour. These data could provide alternative sources for the assessment of phenomena such as antenatal depression, which need to be explored (and validated if proven to be useful). Moreover, the recent increase in the availability of wearable devices could also complement these assessments.

Moreover, this project assessed the role of mobile apps for collecting patient self-reported data through surveys. However, mobile apps can also be used for delivering interventions, and for enhancing the communication between patients and their healthcare providers. Further work should assess how these capabilities could support additional aspects of antenatal mental health care pathways (i.e., diagnostic and therapeutic).

This project established the feasibility of using apps for implementing antenatal depression screening in antenatal clinics and naturalistic settings. However, these apps were not fully integrated with existing clinical systems. This is a crucial aspect that needs to be addressed before a complete evaluation of the cost-effectiveness of app-based antenatal depression screening can be performed. In addition, diagnostic procedures need to be performed in order to rule out potential delivery mode effects. Furthermore, future evaluations should include formal qualitative analyses of stakeholders' perceptions and experiences of using mobile apps for antenatal depression screening.

Although antenatal depression is an important problem, it represents only one aspect of perinatal mental health. Anxiety is another important condition that should be considered. In addition to screening scales, mobile apps can also support the assessment of risk factors. Moreover, perinatal mental health problems can be experienced up to 4 years after childbirth. Therefore, it is important to consider how mobile apps could support the assessment of perinatal mental health at any point during the antenatal and postnatal periods. This would require engaging services other than antenatal care: health visitors, GPs and paediatricians. Furthermore, perinatal mental health problems are not always experienced in isolation. Partners, relatives, friends and children could be affected either through personal experience of mental health problems, or as carers. The role of mobile technology for supporting these social and family networks (when appropriate) should also be explored.

Lastly, it is crucial to determine which subgroups of pregnant women would be best served by the approaches discussed in this project. This would require the recruitment of a more diverse sample of participants, including women with current or recent experience of mental health problems (which were excluded from this project), as well as those with known risk factors for poor antenatal mental health. This might require approaching women who are not engaged with antenatal services.

7.7. Conclusions

This project showed that it is feasible to use mobile apps for the screening and monitoring of antenatal depression. However, the lessons learned could be applied to other health conditions. The design and development of technology for use in clinical contexts need to be sensitive to the setting in which it will be deployed, and to the needs of its users. The latter include both patients and practitioners, which often have different expectations from mobile systems and different motivations for using them. The assessment of apps should consider the impact that this delivery mode can have on the attributes of clinical data, as well as on clinical workflows and care pathways. These assessments should go beyond overall comparisons between apps and traditional delivery modes, and explore the impact that survey design and app features could have on data quality. From a clinical perspective, it is important to explore how mobile technology could lead to improvements in prediction models of disease, and how it could challenge current approaches to healthcare provision. Overall, mobile technology and other forms of technology should be viewed as powerful tools for supporting and complementing the role of healthcare practitioners, empowering patients and enabling the reconfiguration of health services to better meet the needs of their users.

References (Chapter 7)

1. Marcano Belisario JS, Jamsek J, Huckvale K, O'Donoghue J, Morrison CP, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. *Cochrane Database of Systematic Reviews*. 2015; 7:MR000042. Available from: doi: 10.1002/14651858.MR000042.pub2 [Accessed 10 October 2018].
2. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Morrison C, Car J. Tablet computers for implementing NICE antenatal mental health guidelines: protocol of a feasibility study. *BMJ Open*. 2016;6:e009930. Available from: doi: 10.1136/bmjopen-2015-009930 [Accessed 31 July 2018].
3. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Ramchandani P, Morrison C, Car J. Implementation of depression screening in antenatal clinics through tablet computers: results of a feasibility study. *BMC Medical Informatics and Decision Making*. 2017;17:59. Available from: doi: 10.1186/s12911-017-0459-8 [Accessed 31 July 2018].
4. Marcano-Belisario JS, Doherty K, O'Donoghue J, Ramchandani P, Majeed A, Doherty G, Morrison C, Car J. A bespoke mobile application for the longitudinal assessment of depression and mood during pregnancy: protocol of a feasibility study. *BMJ Open*. 2017;7:e014469. Available from: doi: 10.1136/bmjopen-2016-014469 [Accessed 01 August 2018].
5. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey Methodology*. 2nd Edition. New Jersey, US: Wiley; 2009.
6. Lavrakas, PJ (ed.). *Encyclopedia of Survey Research Methods*. London, UK: SAGE Publications; 2008.
7. Fan W, Yan Z. Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behaviour*. 2010;26:132-139. Available from: doi: 10.1016/j.chb.2009.10.015 [Accessed 31 July 2018].
8. Mavletova A, Couper MC. Mobile Web Survey Design: Scrolling versus Paging, SMS versus E-Mail Invitations. *Journal of Survey Statistics and Methodology*. 2014;2(4):498-518. Available from: doi: 10.1093/jssam/smu015 [Accessed 31 July 2018].

9. Conrad FG, Schober MF, Jans M, Orlowski RA, Nielsen D, Levenstein R. Comprehension and engagement in survey interviews with virtual agents. *Frontiers in Psychology*. 2015;6:1578. Available from: doi: 10.3389/fpsyg.2015.01578 [Accessed 31 July 2018].
10. Mavletova A, Couper MC. Grouping of Items in Mobile Web Questionnaires. *Fields Methods*. 2015;28(2):170-193. Available from: doi: 10.1177/1525822X15595151 [Accessed 31 July 2018].
11. Maloshonok N, Terentev E. The impact of visual design and response formats on data quality in a web survey of MOOC students. *Computers in Human Behavior*. 2016;62:506-515. Available from: doi: 10.1016/j.chb.2016.04.025 [Accessed 31 July 2018].
12. Petrovčič A, Petrič G, Manfreda KL. The effect of email invitation elements on response rate in a web survey within an online community. *Computers in Human Behavior*. 2016;56:320-329. Available from: doi: 10.1016/j.chb.2015.11.025 [Accessed 31 July 2018].
13. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, Lenderking WR, Cella D, Basch E, ISPOR ePRO Task Force. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. *Value in Health*. 2009;12(4): 419-29. Available from: doi: 10.1111/j.1524-4733.2008.00470.x [Accessed 31 July 2018].
14. Moskowitz DS, Young SN. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience*. 2006;31(1): 13-20.
15. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges. *Journal of Medical Internet Research*. 2017;19(3):e77. Available from: doi: 10.2196/jmir.7138 [Accessed 05 January 2019].
16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. *Annual Review of Clinical Psychology*. 2008;4:1-32.

17. Schwartz JE, Stone AA. Strategies for analyzing ecological momentary assessment data. *Health Psychology*. 1998;17(1):6-16.

18. Shiffman S. Conceptualizing analyses of ecological momentary assessment data. *Nicotine & Tobacco Research*. 2014;16(Suppl 2):S76-87. Available from: doi: 10.1093/ntr/ntt195 [Accessed 12 May 2017].

19. Groves RM, Lyberg L. Total Survey Error: Past, Present, Future. *Public Opinion Quarterly*. 2010;74(5):849-879. Available from: doi: 10.1093/poq/nfq065 [Accessed 31 July 2018].

20. West P, Van Kleek M, Giordano R, Weal MJ, Shadbolt N. Common Barriers to the Use of Patient-Generated Data Across Clinical Settings. *Proceedins of the 2018 Conference on Human Factors in Computing Systems (CHI),* 21 – 26 April 2018, Motreal, Canada. Available from: doi: 10.1145/3173574.3174058 [Accessed 31 July 2018].

21. Torous J, Baker JT. Why Psychiatry Needs Data Science and Data Science Needs Psychiatry: Connecting With Technology. *JAMA Psychiatry*. 2016;73(1): 3-4. Available from: doi: 10.1001/jamapsychiatry.2015.2622 [Accessed 05 January 2019].

22. Neff G, Nafus D. *Self-Tracking*. London, UK: The MIT Press; 2016.

23. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, Merchant GC, Naughton F, Blandford A. Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *American Journal of Preventive Medicine*. 2016;51(5):833-842. Available from: doi: 10.1016/j.amepre.2016.06.015 [Accessed 31 July 2018].

24. Ahmed R, Aslani P. What is patient adherence? A terminology overview. *International Journal of Clinical Pharmacology*. 2014;36:4-7.

25. Torous J, Nicholas J, Larsen ME, Firth J, Christensen. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence Based Mental Health*. 2018;21(3):116-119. Available from: doi: 10.1136/eb-2018-102891 [Accessed 01 August 2018].

26. Lupton D. Digital health now and in the future: Findings from a participatory design stakeholder workshop. *Digital Health*. 2017;3: 1 – 17. Available from: doi: 10.1177/2055207617740018 [Accessed 22 October 2018].

27. Accelerated Access: Review of innovative medicines and medical technologies supported by Wellcome Trust (2016). *Accelerated Access Review: Final Report*. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/565072/AAR_final.pdf [Accessed 06 January 2019].

28. NHS England (2014). *Five Year Forward View*. Available from: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf [Accessed 06 January 2019].

29. Imison C, Castle-Clarke S, Watson R, Edwards N (2016). *Delivering the benefits of digital health care*. [Online] Nuffield Trust – Evidence for better health care. Available from: https://www.nuffieldtrust.org.uk/research/delivering-the-benefits-of-digital-health-care [Accessed 06 January 2019].

30. Doherty K, Barry M, Marcano-Belisario JS, Arnaud B, Morrison C, Car J, Doherty G. A Mobile App for the Self-Report of Psychological Well-Being During Pregnancy (BrightSelf): Qualitative Design Study. *JMIR Mental Health*. 2018;5(4): e10007. Available from: doi: 10.2196/10007 [Accessed 30 December 2018].

31. Morrison C, Doherty G. Analyzing Engagement with a Web-Based Intervention Platform Through Visualizing Log-Data. *Journal of Medical Internet Research*. 2014;16(11):e252. Available from: doi: 10.2196/jmir.3575 [Accessed 01 August 2018].

32. Ashall-Payne L, Leigh S. *Health Apps, the Good, the Bad, and the Ugly: The Importance of Information*. [Online] NIHR Blog. Available from: https://www.nihr.ac.uk/blogs/health-apps-the-good-the-bad-and-the-ugly-the-importance-of-information/4702 [Accessed 01 August 2018].

33. Patrick K, Hekler EB, Estrin D, Mohr DC, Riper H, Crane D, Godino J, Riley WT. The Pace of Technologic Change: Implications for Digital Health Behavior Intervention Research. *American Journal of Preventive Medicine*. 2016;51(5):816-824. Available from: doi: 10.1016/j.amepre.2016.05.001 [Accessed 01 August 2018].

List of articles published during doctoral studies

Marcano Belisario JS, Huckvale K, Greenfield G, Car J, Gunn L. Smartphone and tablet self management apps for asthma. *Cochrane Database of Systematic Reviews.* 2013;27(11): CD010013. Available from: doi: 10.1002/14651858.CD010013.pub2.

Marcano-Belisario JS, Gupta AK, O'Donoghue J, Morrison C, Car J. Tablet computers for implementing NICE antenatal mental health guidelines: protocol of a feasibility study. *BMJ Open*. 2016;6(1): e009930. Available from: doi: 10.1136/bmjopen-2015-009930.

Barry M, Doherty K, Marcano-Belisario JM, Car J, Morrison C, Doherty G. mHealth for Maternal Mental Health: Everyday Wisdom in Ethical Design. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6 – 11 May, 2017, Denver, Colorado, United States of America. Available from: doi: 10.1145/3025453.3025918.

Marcano-Belisario JS, Gupta AK, O'Donoghue J, Ramchandani P, Morrison C, Car J. Implementation of depression screening in antenatal clinics through tablet computers: results of a feasibility study. *BMC Medical Informatics and Decision Making*. 2017;17:59. Available from: doi: 10.1186/s12911-017-0459-8.

Marcano-Belisario JS, Doherty K, O'Donoghue J, Ramchandani P, Majeed A, Doherty G, Morrison C, Car J. A bespoke mobile application for the longitudinal assessment of depression during pregnancy: protocol of a feasibility study. *BMJ Open*. 2017;7:e014469. Available from: doi: 10.1136/bmjopen-2016-014469.

Doherty K, Barry M, Marcano-Belisario JS, Arnaud B, Morrison C, Car J, Doherty G. A Mobile App for the Self-Report of Psychological Well-Being During Pregnancy (BrightSelf): Qualitative Design Study. *JMIR Mental Health*. 2018;5(4): e10007. Available from: doi: 10.2196/10007.

Doherty K, Marcano Belisario J, Cohn M, Mastellos N, Morrison C, Car J, Doherty G. Engagement with Mental Health Screening on Mobile Devices: Results from an Antenatal Feasibility Study. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 4 – 9 May, 2019, Glasgow, United Kingdom [Accepted for Publication]

Appendix 1. Search strategy for use in Medline through OvidSP

1. exp Data Collection/
2. exp Self-Assessment/
3. exp Health Status/
4. exp Medical Records Systems, Computerized/
5. data.mp.
6. information.mp.
7. diary.mp.
8. 5 or 6 or 7
9. acqui$.mp.
10. gain$.mp.
11. collect$.mp.
12. obtain$.mp.
13. gather$.mp.
14. captu$.mp.
15. entr$.mp.
16. keep$.mp.
17. input$.mp.
18. 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17
19. 8 adj3 18
20. 1 or 2 or 3 or 4 or 19
21. exp Cellular Phone/
22. Computers, Handheld/
23. (handheld or hand-held) adj1 (computer? Or pc?).mp.
24. cell$ phone.mp.
25. mobile phone?.mp.
26. smartphone?.mp.
27. smart-phone.mp.
28. (“personal digital assistant” or PDA).mp.
29. “palmtop computer?”.mp.
30. (tablet adj3 (device? or comput$)).mp.

31. Blackberry.mp.
32. Nokia.mp.
33. Symbian.mp.
34. (windows adj3 (mobile? Or phone?)).mp.
35. INQ.mp.
36. HTC.mp.
37. sidekick.mp.
38. Android.mp.
39. iPhone?.mp.
40. iPad?.mp.
41. Samsung.mp.
42. 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41
43. 20 and 42
44. Limit 43 to yr=”2007 – Current”

Appendix 2. Search strategy for use in EMBASE through OvidSP

1. exp information processing/
2. exp self evaluation/
3. exp health status/
4. exp electronic medical record/
5. data.mp.
6. information.mp.
7. exp self report/
8. exp questionnaire/
9. diary.mp.
10. 5 or 6 or 9
11. acqui$.mp.
12. gain$.mp.
13. collect$.mp.
14. obtain$.mp.
15. gather$.mp.
16. captu$.mp.
17. entr$.mp.
18. keep$.mp.
19. input$.mp.
20. 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19
21. 10 adj3 20
22. 1 or 2 or 3 or 4 or 7 or 8 or 21
23. exp mobile phone/
24. exp microcomputer/
25. (handheld or hand-held) adj1 (computer? Or pc?).mp.
26. cell$ phone.mp.
27. “mobile phone?”.mp.
28. smartphone?.mp.
29. smart-phone.mp.
30. (“personal digital assistant” or PDA).mp.

31. exp personal digital assistant/
32. “palmtop computer?”.mp.
33. (tablet adj3 (device? or comput$)).mp.
34. Blackberry.mp.
35. Nokia.mp.
36. Symbian.mp.
37. (windows adj3 (mobile? Or phone?)).mp.
38. INQ.mp.
39. HTC.mp.
40. sidekick.mp.
41. Android.mp.
42. iPhone?.mp.
43. iPad?.mp.
44. Samsung.mp.
45. 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44
46. 22 and 45
47. Limit 46 to yr=”2007 – Current”

Appendix 3. Search strategy for use in PsycINFO through OvidSP

1. exp Data Collection/
2. exp Self Evaluation/
3. exp “Quality of Life”/
4. exp Questionnaires/
5. exp Psychometrics/
6. exp Medical Records/
7. exp Surveys/
8. data.mp.
9. exp Information/
10. information.mp.
11. diary.mp.
12. 8 or 10 or 11
13. acqui$.mp.
14. gain$.mp.
15. collect$.mp.
16. obtain$.mp.
17. gather$.mp.
18. captu$.mp.
19. entr$.mp.
20. keep$.mp.
21. input$.mp.
22. 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21
23. 12 adj3 22
24. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 9 or 23
25. exp Cellular Phone/
26. exp Mobile Devices/
27. (handheld or hand-held) adj1 (computer? Or pc?).mp.
28. cell$ phone?.mp.
29. “mobile phone?”.mp.
30. smartphone?.mp.

31. smart-phone?.mp.
32. (“personal digital assistant” or PDA).mp.
33. “palmtop computer?”.mp.
34. (tablet adj3 (device? or comput$)).mp.
35. Blackberry.mp.
36. Nokia.mp.
37. Symbian.mp.
38. (windows adj3 (mobile? Or phone?)).mp.
39. INQ.mp.
40. HTC.mp.
41. sidekick.mp.
42. Android.mp.
43. iPhone?.mp.
44. iPad?.mp.
45. Samsung.mp.
46. 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45
47. 24 and 46
48. Limit 47 to yr=”2007 – Current”

Appendix 4. Search strategy for use in IEEEXplore

(((data collection) OR (data entry) OR (data gathering) OR (questionnaires) OR (self assessment) OR (self evaluation) OR (diary) OR (data keeping) OR (psychometrics) OR (data capture) OR (quality of life)) AND ((smartphone) OR (smart-phone) OR (handheld computer) OR (mobile phone) OR (cellular phone) OR (cell phone) OR (mobile device) OR (tablet) OR (tablet computer) OR (tablet device) OR (iPhone) OR (iPad) OR (Samsung) OR (Nokia) OR (Windows Phone) OR (Blackberry) OR (HTC) OR (INQ) OR (Android)))

Appendix 5. Search strategy for use in Web of Science

1. TS=(data collection)
2. TS=(data capture)
3. TS=(self assessment)
4. TS=(self report)
5. TS=(questionnaire)
6. TS=(data entry)
7. TS=(data gathering)
8. TS=(diary)
9. TS=(psychometrics)
10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9
11. TS=(mobile phone)
12. TS=(mobile device)
13. TS=(cell phone)
14. TS=(cellular phone)
15. TS=(smartphone)
16. TS=(smart-phone)
17. TS=(handheld computer)
18. TS=(handheld device)
19. TS=(hand-held computer)
20. TS=(hand-held device)
21. TS=(personal digital assistant)
22. TS=(PDA)
23. TS=(tablet)
24. TS=(tablet device)
25. TS=(tablet computer)
26. TS=(iPhone)
27. TS=(iPad)
28. TS=(Samsung)
29. TS=(palmtop computer)
30. TS=(Nokia)

31. TS=(Blackberry)
32. TS=(Android)
33. TS=(HTC)
34. TS=(INQ)
35. TS=(Windows phone)
36. TS=(Sidekick)
37. 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36
38. 10 and 37

Appendix 6. Search strategy for use in CABI: CAB Abstracts

1. TS=(data collection)
2. TS=(data capture)
3. TS=(self assessment)
4. TS=(self report)
5. TS=(questionnaire)
6. TS=(data entry)
7. TS=(data gathering)
8. TS=(diary)
9. TS=(psychometrics)
10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9
11. TS=(mobile phone)
12. TS=(mobile device)
13. TS=(cell phone)
14. TS=(cellular phone)
15. TS=(smartphone)
16. TS=(smart-phone)
17. TS=(handheld computer)
18. TS=(handheld device)
19. TS=(hand-held computer)
20. TS=(hand-held device)
21. TS=(personal digital assistant)
22. TS=(PDA)
23. TS=(tablet)
24. TS=(tablet device)
25. TS=(tablet computer)
26. TS=(iPhone)
27. TS=(iPad)
28. TS=(Samsung)
29. TS=(palmtop computer)
30. TS=(Nokia)

31. TS=(Blackberry)
32. TS=(Android)
33. TS=(HTC)
34. TS=(INQ)
35. TS=(Windows phone)
36. TS=(Sidekick)
37. 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36
38. 10 and 37

Appendix 7. Search strategy for use in Current Contents Connect

1. TS=(data collection)
2. TS=(data capture)
3. TS=(self assessment)
4. TS=(self report)
5. TS=(questionnaire)
6. TS=(data entry)
7. TS=(data gathering)
8. TS=(diary)
9. TS=(psychometrics)
10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9
11. TS=(mobile phone)
12. TS=(mobile device)
13. TS=(cell phone)
14. TS=(cellular phone)
15. TS=(smartphone)
16. TS=(smart-phone)
17. TS=(handheld computer)
18. TS=(handheld device)
19. TS=(hand-held computer)
20. TS=(hand-held device)
21. TS=(personal digital assistant)
22. TS=(PDA)
23. TS=(tablet)
24. TS=(tablet device)
25. TS=(tablet computer)
26. TS=(iPhone)
27. TS=(iPad)
28. TS=(Samsung)
29. TS=(palmtop computer)
30. TS=(Nokia)

31. TS=(Blackberry)
32. TS=(Android)
33. TS=(HTC)
34. TS=(INQ)
35. TS=(Windows phone)
36. TS=(Sidekick)
37. 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36
38. 10 and 37

Appendix 8. Search strategy for use in ACM Digital

("data collection") and ("mobile phone" or "smartphone" and "smart-phone" and "cell phone" or "cellular phone" or tablet or "tablet computer" or "tablet device")

Appendix 9. Search strategy for use in ERIC

(IF(data collection) or IF(self evaluation) or IF(self assessment) or IF(questionnaire) or IF(psychometrics) or IF(data entry) or IF(data capture) or IF(diary) or IF(data gathering) or IF(information gathering) or IF(quality of life)) and (IF(mobile phone) or IF(cell phone) or IF(cellular phone) or IF(smartphone) or IF(smart-phone) IF(handheld computer) or IF(palmtop) or IF(iPhone) or IF(iPad) or IF(Samsung) or IF(Windows phone) IF(Blackberry) or IF(Nokia) or IF(HTC) or IF(Symbian) or IF(Android) or IF(Sidekick) or IF(INQ) or IF(tablet) or IF(tablet computer) or IF(tablet device))

Appendix 10. Search strategy for use in Sociological Abstracts

(IF(data collection) or IF(self evaluation) or IF(self assessment) or IF(questionnaire) or IF(psychometrics) or IF(data entry) or IF(data capture) or IF(diary) or IF(data gathering) or IF(information gathering) or IF(quality of life)) and (IF(mobile phone) or IF(cell phone) or IF(cellular phone) or IF(smartphone) or IF(smart-phone) IF(handheld computer) or IF(palmtop) or IF(iPhone) or IF(iPad) or IF(Samsung) or IF(Windows phone) IF(Blackberry) or IF(Nokia) or IF(HTC) or IF(Symbian) or IF(Android) or IF(Sidekick) or IF(INQ) or IF(tablet) or IF(tablet computer) or IF(tablet device))

Appendix 11. Search strategy Health Management Information Consortium

1. exp Data Collection/
2. exp Self Assessment/
3. exp health status/
4. exp Surveys/
5. exp Health surveys/
6. exp Questionnaires
7. data.mp.
8. information.mp.
9. diary.mp.
10. 7 or 8 or 9
11. acqui$.mp.
12. gain$.mp.
13. collect$.mp.
14. obtain$.mp.
15. gather$.mp.
16. captu$.mp.
17. entr$.mp.
18. keep$.mp.
19. input$.mp.
20. 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19
21. 10 adj3 20
22. 1 or 2 or 3 or 4 or 5 or 6 or 21
23. exp Mobile telephones/
24. (handheld or hand-held) adj1 (computer? Or pc?).mp.
25. cell$ phone.mp.
26. “mobile phone?”.mp.
27. smartphone?.mp.
28. smart-phone.mp.

29. (“personal digital assistant” or PDA).mp.
30. “palmtop computer?”.mp.
31. (tablet adj3 (device? or comput$)).mp.
32. Blackberry.mp.
33. Nokia.mp.
34. Symbian.mp.
35. (windows adj3 (mobile? Or phone?)).mp.
36. HTC.mp.
37. Android.mp.
38. iPhone?.mp.
39. iPad?.mp.
40. 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39
41. 22 and 40
42. Limit 41 to yr=”2007 – Current”

Appendix 12. Search strategy for use in CENTRAL

1. MeSh descriptor: [Data Collection] explode all trees
2. MeSH descriptor: [Self-Assessment] explode all trees
3. MeSh descriptor: [Health Status] explode all tress
4. “data”.ti,ab,kw
5. “information”.ti,ab,kw
6. “diary”.ti,ab,kw
7. 5 or 6
8. “acquisition”.ti,ab,kw
9. “gain”.ti,ab,kw
10. “collection”.ti,ab,kw
11. “obtain”.ti,ab,kw
12. “gather”.ti,ab,kw
13. “capture”.ti,ab,kw
14. “entry”.ti,ab,kw
15. “keep”.ti,ab,kw
16. “input”.ti,ab,kw
17. 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16
18. 7 adj3 17
19. 1 or 2 or 3 or 18
20. MeSH descriptor: [Cellular Phone] explode all trees
21. MeSH descriptor: [Computers, Handheld] explode all trees
22. “cell phone”.ti,ab,kw
23. “mobile phone”.ti,ab,kw
24. “smartphone”.ti,ab,kw
25. “personal digital assistant”.ti,ab,kw
26. “palmtop”.ti,ab,kw
27. “tablet computer”.ti,ab,kw
28. “blackberry”.ti,ab,kw
29. “Nokia”.ti,ab,kw
30. “HTC”.ti,ab,kw

31. “Android”.ti,ab,kw
32. “iPhone”.ti,ab,kw
33. “iPad”.ti,ab,kw
34. “Samsung”.ti,ab,kw
35. 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34
36. 19 and 35

Appendix 13. Search strategy for use in ClinicalTrials.gov

(questionnaires OR surveys) AND (smartphones OR mobile OR phone OR apps)

Appendix 14. Search strategy for us in ProQuest Dissertation and Theses

(IF(data collection) or IF(self evaluation) or IF(self assessment) or IF(questionnaire) or IF(psychometrics) or IF(data entry) or IF(data capture) or IF(diary) or IF(data gathering) or IF(information gathering) or IF(quality of life)) and (IF(mobile phone) or IF(cell phone) or IF(cellular phone) or IF(smartphone) or IF(smart-phone) IF(handheld computer) or IF(palmtop) or IF(iPhone) or IF(iPad) or IF(Samsung) or IF(Windows phone) IF(Blackberry) or IF(Nokia) or IF(HTC) or IF(Symbian) or IF(Android) or IF(Sidekick) or IF(INQ) or IF(tablet) or IF(tablet computer) or IF(tablet device))

Appendix 15. Initial Search Results

Electronic Bibliographic Database	Dates searched	Results Before De-Duplication	Results After De-Duplication
MEDLINE	Jan 2007 – Jun 2014	2700	2579
EMBASE	Jan 2007 – Jun 2014	7701	7491
PsycINFO	Jan 2007 – Jun 2014	664	663
IEEEXplore	Jan 2007 – Jun 2014	2476	2472
Web of Science	Jan 2007 – Jun 2014	4555	4546
CABI: CAB Abstracts	Jan 2007 – Jun 2014	867	863
Current Contents Connect	Jan 2007 – Jun 2014	2868	2864
ACM Digital Library	Jan 2007 – Jun 2014	2306	104
ERIC	Jan 2007 – Jun 2014	5	5
Sociological Abstracts	Jan 2007 – Jun 2014	26	26
Campbell Library	Jan 2007 – Jun 2014	0	0
ClinicalTrials.gov	All	1148	1148
WHO ICTRP	All	2253	2224
OpenGrey	All	0	0
Mobile Active	All	0	0
Dissertation & Theses	Jan 2007 – Jul 2014	27	27
Health Management Information Consortium	Jan 2007 – Jul 2014	37	36
CENTRAL	Jan 2007 – Jul 2014	179	179
Google Scholar	All	410	0
Combined Databases		**28222**	**17168**

Appendix 16. Update search results

Electronic Bibliographic Database	Dates searched	Results Before De-Duplication	Results After De-Duplication
MEDLINE	Jun 2014 – Apr 2015	649	603
EMBASE	Jun 2014 – Apr 2015	2234	2190
PsycINFO	Jun 2014 – Apr 2015	236	236
IEEEXplore	Jun 2014 – Apr 2015	1415	1398
Web of Science	Jun 2014 – Apr 2015	1182	1182
CABI: CAB Abstracts	Jun 2014 – Apr 2015	155	155
Current Contents Connect	Jun 2014 – Apr 2015	887	887
ACM Digital Library	Jun 2014 – Apr 2015	290	0
ERIC	Jun 2014 – Apr 2015	0	0
Sociological Abstracts	Jun 2014 – Apr 2015	0	0
Campbell Library	Jun 2014 – Apr 2015	0	0
ClinicalTrials.gov	Jun 2014 – Apr 2015	396	396
WHO ICTRP	Jun 2014 – Apr 2015	453	453
OpenGrey	Jun 2014 – Apr 2015	0	0
MobileActive	Jun 2014 – Apr 2015	0	0
Dissertation & Theses	Jun 2014 – Apr 2015	3	3
Health Management Information Consortium	Jun 2014 – Apr 2015	6	6
CENTRAL	Jun 2014 – Apr 2015	9	9
Google Scholar	All	0	0
Combined Databases		**7065**	**5507**

Appendix 17. Characteristics of Excluded Studies

Study ID	Reason for exclusion
Abernethy 2008	Study conducted before 2007
Ahmad 2012	No suitable comparator(s)
Aktas 2014	No suitable comparator(s) or study design
Alhajji 2009	No suitable comparator(s)
Allena 2012	The intervention and the study design did not meet the inclusion criteria. Device used was released before 2007
Alsip 2014	The intervention, comparison, outcomes and study design did not meet the inclusion criteria
Bakshi 2013a	The comparison, outcomes and study design did not meet the inclusion criteria
Bakshi 2013b	The intervention, comparison, outcomes and study design did not meet the inclusion criteria
Barentsz 2014	Study design not suitable
Bartlett 2013a	No suitable comparator(s)
Bartlett 2013b	No suitable comparator(s)
Beasley 2008	The intervention did not meet the inclusion criteria
Bellamy 2009a	The intervention did not meet the inclusion criteria
Bellamy 2009b	The intervention did not meet the inclusion criteria
Bellamy 2011a	The intervention did not meet the inclusion criteria
Bellamy 2011b	No suitable comparator(s)
Ben-Zeev 2012	The intervention, comparison and study design did not meet the inclusion criteria
Bernabe-Ortiz 2008	The intervention and the study design did not meet the inclusion criteria
Bernhardt 2009	The intervention, comparison and outcomes did not meet the inclusion criteria
Berry 2014	The intervention, comparison, and outcomes did not meet the inclusion criteria
Bethoux 2014	No suitable comparator(s)
Bexelius 2010	The outcomes and the study design did not meet the inclusion criteria
Blaivas 2013a	The comparison, outcomes and study design did not meet the inclusion criteria

Blaivas 2013b	The comparison, outcomes and study design did not meet the inclusion criteria
Blum 2014	The intervention, comparison, and study design did not meet the inclusion criteria
Bockenek 2014	The participants, intervention, and comparison did not meet the inclusion criteria
Bokhour 2013	The intervention and outcomes did not meet the inclusion criteria
Bond 2013	No suitable comparator(s) or study design
Boushey 2009	The intervention, outcomes and study design did not meet the inclusion criteria
Bradbury 2012	The intervention, comparison and study design did not meet the inclusion criteria
Braun 2008	The participants and the study design did not meet the inclusion criteria
Burke 2009	The intervention did not meet the inclusion criteria
Buskirk 2014	The intervention did not meet the inclusion criteria
Carter 2013a	No suitable comparator(s) or study design
Carter 2013b	The comparison did not meet the inclusion criteria
Christie 2013	The intervention did not meet the inclusion criteria
Clionsky 2014	The intervention, and comparison did not meet the inclusion criteria
Cook 2007	The intervention did not meet the inclusion criteria
Croff 2012	The participants, intervention and outcomes did not meet the inclusion criteria
Cudlip 2014	The comparison, outcomes and study design did not meet the inclusion criteria
Cunningham 2013	The intervention, comparison and study design did not meet the inclusion criteria
Dale 2007	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
de Bruijne 2013	The intervention did not meet the inclusion criteria
DeMaria 2012	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Denny 2008	No suitable outcome(s) or study design
Depp 2012	Study excluded during data extraction: responses were entered via a web browser;

	surveys were delivered via a SMS that automatically redirected participants to the URL of the survey.
Desai 2012	The intervention, comparison and outcomes did not meet the inclusion criteria
Dewit 2012	Study design not suitable
Duncan 2012	The intervention and comparison did not meet the inclusion criteria
Dupont 2009	Study design not suitable
Dy 2012	Study design not suitable
Edwards 2008	The intervention did not meet the inclusion criteria
Escandon 2008	The intervention and the study design did not meet the inclusion criteria
Eskenazi 2014	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Fanning 2014	Study excluded during data extraction: responses were entered via a web browser.
Farach 2013	No suitable comparator(s) or study design
Faurholt-Jepsen 2013	The intervention, comparison and outcomes did not meet the inclusion criteria
Fritz 2012	Study design not suitable
Galliber 2008	The participants, intervention and comparison did not meet the inclusion criteria
Garcia 2010	The intervention, comparison and outcomes did not meet the inclusion criteria
Gibbons 2011	The intervention did not meet the inclusion criteria
Giesinger 2013	The intervention, comparison and study design did not meet the inclusion criteria
Glaser 2013	The participants, outcomes and study design did not meet the inclusion criteria
Goldstein 2011	Study design not suitable
Gupta 2013	The participants, intervention and study design did not meet the inclusion criteria
Gurland 2010a	The comparison, outcomes and study design did not meet the inclusion criteria
Gurland 2010b	The intervention, comparison and study design did not meet the inclusion criteria
Hallum-Montes 2013	No suitable comparator(s) or outcome(s)

Harralson 2013	The intervention, comparison, outcomes and study design did not meet the inclusion criteria
Harris 2013	No suitable comparator(s) or study design
Hashemian 2012	The participants, comparison, outcomes and study design did not meet the inclusion criteria
Haver 2011	Study excluded during data extraction: non-health survey
Heiberg 2007	The intervention did not meet the inclusion criteria
Hollen 2013	Study design not suitable
Huang 2010	The intervention did not meet the inclusion criteria
Huang 2012	The intervention, comparison, outcomes and study design did not meet the inclusion criteria
Huguet 2014	The intervention, comparison, outcomes and study design did not meet the inclusion criteria
Hundeshagen 2013	The participants, comparison, outcomes and study design did not meet the inclusion criteria
Hutchesson 2015	The intervention did not meet the inclusion criteria
Isara 2013	The intervention, comparison and study design did not meet the inclusion criteria
Jacob 2012	Study design not suitable
Jaspan 2007	The intervention did not meet the inclusion criteria
Johnson 2014	The intervention, and the study design did not meet the inclusion criteria
Juniper 2009	The intervention did not meet the inclusion criteria
Junker 2008	The intervention did not meet the inclusion criteria
Kajander 2007	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Kauer 2012	No suitable comparator(s) or outcome(s)
Kaufman 2013	No suitable comparator(s) or study design
Kelly 2014	No suitable comparator(s), study design or outcome(s)
Khair 2014a	Study design not suitable
Khair 2014b	Study design not suitable
Khor 2014a	No suitable comparator(s) or outcome(s)
Khor 2014b	No suitable comparator(s) or outcome(s)

Khraishi 2013	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Kimel 2010	The intervention and the study design did not meet the inclusion criteria
King 2013	The participants and study design did not meet the inclusion criteria
Kirwan 2012	No suitable comparator(s) or study design
Kochan 2007	The intervention did not meet the inclusion criteria
Krogh 2013	The intervention did not meet the inclusion criteria
Kuntsche 2013	No suitable comparator(s) or outcome(s)
Kuntsche 2014	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Lam 2010	The intervention did not meet the inclusion criteria
Lange 2014	No suitable comparator(s) or study design
Lee 2010	The participants, intervention and study design did not meet the inclusion criteria
Lee 2014	The intervention, comparison and study design did not meet the inclusion criteria
Levine 2012	The intervention, comparison, outcomes and study design did not meet the inclusion criteria
Lundy 2013	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Malotte 2011	No suitable comparator(s) or study design
Mangera 2014	The intervention did not meet the inclusion criteria
Marceau 2007	The intervention did not meet the inclusion criteria
Marceau 2010	The intervention did not meet the inclusion criteria
Martin 2012	No suitable comparator(s), outcome(s) or study design
Matthew 2007a	The intervention did not meet the inclusion criteria
Matthew 2007b	The intervention did not meet the inclusion criteria
Mavletova 2013	Study excluded during data extraction: responses were entered via a mobile-enabled web browser; both smartphones and feature phones were used to collect responses

Mays 2010	The intervention did not meet the inclusion criteria
McCaw 2010	The intervention did not meet the inclusion criteria
McIntosh 2013	Study design not suitable
Michalak 2009	The intervention did not meet the inclusion criteria
Miller 2013	No suitable comparator(s), outcome(s) or study design
Mulvaney 2012	The intervention, comparison and outcomes did not meet the inclusion criteria
Nishiguchi 2014	No suitable comparator(s) or study design
Oliver 2013	The intervention did not meet the inclusion criteria
Pakhare 2013	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Patel 2012	No suitable comparator(s) or outcome(s)
Patnaik 2009	The participants and study design did not meet the inclusion criteria
Pau 2013	No suitable comparator(s) or study design
Pfaeffli 2013	No suitable comparator(s)
Phillips 2014	No suitable comparator(s), outcome(s) or study design
Polak 2014	The intervention did not meet the inclusion criteria
Quadri 2012	Study design not suitable
Rao 2014	No suitable comparator
Raptis 2011	Participants and outcomes did not meet the inclusion criteria
Richter 2008	The intervention did not meet the inclusion criteria
Ring 2008	The intervention did not meet the inclusion criteria
Roth 2014	No suitable comparator(s) or study design
Runyan 2013	The intervention, comparison and outcomes did not meet the inclusion criteria
Russman 2014	No suitable comparator(s) or study design
Sage 2012	The study design did not meet the inclusion criteria
Sander 2012	No suitable comparator(s)
Scheers 2012	The intervention, comparison and study design did not meet the inclusion criteria
Schlechtweg 2013	Study design not suitable

Seebregts 2009	The intervention did not meet the inclusion criteria
Shafran 2009	No suitable comparator(s), outcome(s) or study design
Shapiro 2011	The intervention and the outcomes did not meet the inclusion criteria
Shay 2009	The intervention did not meet the inclusion criteria
Short 2013	The participants, intervention and study design did not meet the inclusion criteria
Smith 2011	Study design not suitable
Smith 2014	The intervention did not meet the inclusion criteria
Spark 2015	The intervention did not meet the inclusion criteria
Sternfeld 2012	The intervention did not meet the inclusion criteria
Stukenborg 2013	No suitable comparator(s) or outcome(s)
Swartz 2007	The intervention did not meet the inclusion criteria
Tegang 2009	The participants, intervention, comparison and study design did not meet the inclusion criteria
Temple 2014	Study design not suitable
Tolley 2013	No suitable comparator(s) or study design
Trapl 2007	The intervention did not meet the inclusion criteria
Trapl 2013	The intervention did not meet the inclusion criteria
Tully 2014	No suitable comparator(s) or study design
Tyser 2015	The intervention did not meet the inclusion criteria
Unver 2009	The intervention, comparison and study design did not meet the inclusion criteria
van Duinen 2008	The intervention did not meet the inclusion criteria
van Heerden 2014	No suitable comparator(s) or study design
Vargas 2010	No suitable comparator(s)
Viana 2014	No suitable comparator(s)
Vinney 2012	The intervention did not meet the inclusion criteria
Walther 2011	The participants, intervention, comparison and outcomes did not meet the inclusion criteria
Wells 2014	Study excluded during data extraction: non-health survey
Wharton 2014	No suitable comparator(s)

Wilcox 2012	The participants, intervention, comparison, outcomes and study design did not meet the inclusion criteria
Wilson 2013a	No suitable comparator(s), outcome(s) or study design
Wilson 2013b	No suitable comparator(s), outcome(s) or study design
Witt 2015	Study design not suitable
Wofford 2014	No suitable comparator(s) or study design
Wood 2011	The intervention did not meet the inclusion criteria
Woods 2009	This study was excluded during data extraction: pre-2007 device
Wundes 2011	Study design not suitable
Yon 2007	The intervention and outcomes did not meet the inclusion criteria
Yu 2009	The participants and intervention did not meet the inclusion criteria
Zhang 2012	The participants and intervention did not meet the inclusion criteria
Zhu 2009	No suitable comparator(s) or outcome(s)

Appendix 18. Characteristics of Ainsworth and colleagues (2013)

Methods	*Study design*: crossover trial *Country*: UK *Incentives provided*: £50 worth of phone credit, plus £30 in cash upon completion of both sampling periods *Type of device & platform*: smartphone app designed to run on Android devices; an Orange San Francisco device was used *Functionality*: configurable number of/times questions are displayed on each day; configurable questions; multiple question sets; question branching; questionnaire timeout; time stamping; and skip procedures *Human Computer Interaction*: user-definable alerts delivered at semi-random intervals; snoozing of alerts; one alert per question set; one question per page; navigation through pages of questions; continuous slider bar mapped onto a 7-point Likert scale; automatic saving of responses *Sampling schedule*: 4 times a day for 6 days *Additional interventions*: training; and phone calls during the sampling period
Data	*Name of survey(s)*: diagnostic assessment items assessing 7 symptom dimensions: hopelessness, depression, hallucinations, anxiety, grandiosity, paranoia and delusions *Validation status*: composite instrument derived from previously validated scales *Application of the survey*: mental health assessment *Population*: 24 patients diagnosed with schizophrenia or schizoaffective disorder *Age group*: adults, mean age 33 years old (SD 9.5), range 18 to 49 years *Gender composition of the sample*: 79.17% male participants; 20.83% female participants *Setting of survey completion:* naturalistic
Comparisons	SMS
Outcomes	Equivalence: mean score differences Data completeness: mean number of data entries completed Time to completion: meant time taken to complete the questions Adherence to data collection protocol: proportion of individuals completing at least one third of all possible data points Acceptability: reactivity to the methodology; successful integration with an individual's daily routine; length of time

		participants would be willing to use the delivery mode; ease of use; and preference
Notes		A semi-structured interview (PANSS) was conducted after each sampling period. Participants had only 15 minutes from the initial alert within which they had to complete the questions; this was thought to reduce the likelihood of self selection bias (i.e., participants answering questions only when they were asymptomatic)
		Risk of Bias
Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Low risk	Authors evaluated the interaction between sampling period and method of assessment, and found that order of administration did not significantly predict the total number of entries an individual completed, or the length of time it took to complete each entry
First period data available	Low risk	Data from both sampling periods were included in the statistical analyses
Correct analysis	Low risk	Spearman correlation coefficient
Comparability of results with those from parallel trials	Low risk	Appropriate randomisation procedure was followed; adequate washout period (7 days); and absence of carry-over effect.

Appendix 19. Characteristics of Brunger and colleagues (2015)

Methods	*Study design*: paired repeated measures design *Country*: UK *Incentives provided*: none *Type of device & platform*: app designed to run on an iPad mini (iOS) *Functionality*: none mentioned *Human Computer Interaction*: 5 questions (one question per page) mapped onto a 100mm VAS; 3 questions were end-anchored with *Not at all* and *Extremely*, whereas one question was end-anchored with *Nothing at all* and *Very much* and another question with *Weak* and *Very strong*; respondents selected their answers by using their fingers on the touchscreen; there was a greater distance between the end anchor and the line than in the comparison group; data was stored in the device and automatically transferred into a secure database via a wireless connection *Sampling schedull*: one survey completion session with endpoint assessments at baseline and 0, 30, 60 and 120 minutes after consuming either a low-energy or a high-energy drink *Additional interventions*: none mentioned
Data	*Name of survey*: five appetite rating questions (How hungry are you? How full are you? How satiated are you? How much do you think you could eat right now? and How strong is your desire to eat?) mapped onto a VAS *Validation status*: unclear; questions have been recommended for use in appetite studies *Application of the survey*: appetite ratings *Population*: 18 healthy adults with BMI between 18 and 28 kg/m² *Age group*: adults, mean age 28.5 years old (SD 5.5) *Gender composition of the sample*: 50% male participants; 50% female participants *Setting of survey completion*: clinical

Comparisons	PDA: iPAQ	
Outcomes	Equivalence: correlation coefficient	
Notes	The aim of this study was to validate an improved iPad based rating system relative to an existing iPAQ based system, while contrasting a shorter (64 mm) and a longer (100 mm) scale length.	
Risk of Bias		
Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Low risk	Order of the device was counterbalanced. In addition, the authors accounted for type of device in their statistical analyses
First period data available	Low risk	For each pair of assessments the authors included data from both sampling periods.
Correct Analysis	Low risk	Correlation coefficients
Comparability of results with those from parallel trials	High risk	Although the study authors counterbalanced the order of the device and considered the type of device in their statistical analyses, they recruited a small sample of participants (i.e., 18 participants)

Appendix 20. Characteristics of Bush and colleagues (2013)

Methods	*Study design*: crossover trial *Country*: US *Incentives provided*: none mentioned *Type of device & platform*: MobileScreener, an app running on iPhone devices (iOS) *Functionality*: not specified *Human Computer Interaction*: not specified *Sampling schedule*: one survey completion session taking place on one day *Additional interventions*: none mentioned
Data	*Name of survey*: Mobile Screener, consisting of the PTSD Checklist, PHQ 9, Revised Suicidal Ideation Scale, Deployment Risk and Resilience Inventory-Unit Support, Dimensions of Anger 5, Sleep Evaluation Scale and TBI Self Report of Symptoms *Validation status*: composite instrument made up of validated instruments *Application of the survey*: mental health assessment *Population*: 45 healthy, active military personnel *Age group*: adults *Gender composition of the sample*: 77.78% male participants; and 22.22% female participants *Setting of survey completion*: clinical
Comparisons	Laptop Pen-and-paper
Outcomes	Equivalence: mean score differences and ICC Acceptability: ease of use; likelihood of use; and preference
Notes	One of the objectives of this study was to develop and validate a smartphone app to be used amongst a highly mobile military population. Participants were asked to complete the iPhone measurement for a second time, and this was used to calculate test-retest reliability

Risk of Bias

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)

Carry-over effect	High risk	The presence of carry-over effect was not formally tested; the washout period does not seem adequate (participants completed both sampling periods within 90 minutes).
First period data available	Low risk	Data from both sampling periods were included in the statistical analyses.
Correct Analysis	Low risk	Mean scores and ICC
Comparability of results with those from parallel trials	High risk	Washout period was inadequate

Appendix 21. Characteristics of Garcia-Palacios and colleagues (2014)

Methods	*Study design*: crossover trial *Country*: Spain *Incentives provided*: free psychological sessions for the treatment of fibromyalgia syndrome *Type of device & platform*: F-EMA, an app running on a HTC Diamond 1 smartphone (Windows Mobile OS) *Functionality*: configurable number of questions displayed on each day; and time stamping of data entries *Human Computer Interaction*: audio alerts; configurable alert schedule; reminders every minute during the first 15 minutes after the initial alert and then every 15 minutes during the next hour; audio-recorded instructions *Sampling schedule*: 3 times a day for 7 days *Additional interventions*: none reported
Data	*Name of survey*: EMA measures assessing pain intensity, fatigue intensity, and mood *Validation status*: unclear *Application of the survey*: functional status assessment, pain assessment, and mental health assessment *Population*: 47 patients diagnosed with fibromyalgia syndrome *Age group*: adults, mean age 48.1 years (SD 7.95), range 37 to 65 years *Gender composition of the sample*: all female participants *Setting of survey completion*: naturalistic
Comparisons	Pen-and-Paper
Outcomes	Equivalence: mean score differences Adherence to sampling protocol: mean number of both complete and incomplete records Acceptability: acceptability and preference
Notes	A technological profile questionnaire was developed for this study. Participants were asked to attend the clinic at the end of each sampling period and complete a weekly rating of pain and fatigue
Risk of bias	

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	High risk	The presence of carry-over effect was not formally assessed; the washout period was insufficient (participants attended the clinic after the first sampling period for an assessment, and to switch over to the alternative delivery mode)
First period data available	Low risk	Data from both sampling periods were included in the statistical analyses
Correct Analysis	High risk	Data on mood assessments were not reported; data from 7 participants were excluded as they failed to show up to the assessment session at the end of the first week of sampling
Comparability of results with those from parallel trials	Unclear risk	Randomisation procedure was not specified; presence of carry-over effect was not explored; and unclear whether the washout period was appropriate

Appendix 22. Khraishi and colleagues (2012)

Methods	*Study design*: crossover trial *Country*: Canada *Incentives provided*: none mentioned *Type of device & platform*: app running on an Apple iPad *Functionality*: not reported *Human Computer Interaction*: not reported *Sampling schedule*: one survey completion session taking place on one day *Additional interventions*: not reported
Data	*Name of survey*: Health Assessment Questionnaire *Validation status*: validated *Application of the survey*: functional status assessment *Population*: 32 patients diagnosed with psoriatic arthritis or rheumatoid arthritis *Age group*: adults, range 30 to 70 years *Gender composition of the sample*: 62.5% female participants *Setting of survey completion*: clinical
Comparisons	Pen-and-paper
Outcomes	Equivalence: mean score difference Time to completion: mean time taken to complete the survey questionnaire Acceptability: ease of use; perception of time taken to complete the questionnaire; preference; perceived benefit of the delivery mode
Notes	Type of publication: abstract

Risk of Bias

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Unclear risk	It is unclear if the presence of carry-over effect was tested and there is no information available on the duration of the washout period
First period data available	Unclear risk	It is unclear if data from both sampling periods were included in the statistical analyses

Correct Analysis	Unclear risk	Details of the statistical analyses used are not available
Comparability of results with those from parallel trials	Unclear risk	Every other patient was assigned to the reverse order; it is unclear if authors tested for carry-over effects; no information on the washout period

Appendix 23. Characteristics of Kim and colleagues (2014)

Methods	*Study design*: crossover trial *Country*: Republic of Korea *Incentives provided*: none provided *Type of device & platform*: app running on Android devices *Functionality*: not reported *Human Computer Interaction*: one question per page; navigation through multiple pages; users were allowed to correct previous answers; users had to tap 'Save' in order to save their answers; and transmission of data was automatic *Sampling schedule:* one survey completion session that took place on a single day *Additional interventions*: not reported
Data	*Name of survey*: International Prostate Symptom Score (IPSS) *Validation status*: validated *Application of the survey*: symptom score index *Population*: 1581 patients with lower urinary tract symptoms *Age group*: adults, mean age 58.5 years (SD 7.22), range 40 to 79 years *Gender composition of the sample*: 100% male participants *Setting of survey completion:* clinical
Comparisons	Pen-and-paper
Outcomes	Equivalence: mean score differences; and ICC Acceptability: willingness to use a particular method; preferred method
Notes	N/A

Risk of Bias

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Low risk	The presence of carry-over effect was not formally tested; however, a washout period of one week was chosen to reduce the risk of a carry-over effect

First period data available	Low risk	Data from both sampling periods were included in the statistical analyses
Correct Analysis	Low risk	ICC and a two-way random effect model
Comparability of results with those from parallel trials	Low risk	An effort was made to minimise the potential of a carry-over effect (washout period of one week); the authors managed to recruit a large sample of participants (N = 1581)

Appendix 24. Characteristics of Lamber and colleagues (2012)

Methods	*Study design*: RCT *Country*: Italy *Incentives provided*: none reported *Type of device & platform*: MobiDay, an app running on a Nokia N97 smartphone (Symbian OS) *Functionality*: no description provided *Human Computer Interaction*: one question per page, automatic saving of data, respondents were allowed to suspend their tasks and return to the questionnaire whenever it was convenient for them *Sampling schedule:* one survey completion session taking place in one day *Additional interventions*: none reported
Data	*Name of survey*: European Organization for Research and Treatment of Cancer Quality of Life Questionnaire - C30 (EORTC QLQ-C30) *Validation status*: validated instrument *Application of the survey*: assessment of health-related quality of life in patients diagnosed with cancer *Population*: patients diagnosed with cancer *Age group*: adults, range 30 to 80 years old *Gender composition of the sample*: not reported *Setting of survey completion*: clinical
Comparisons	Laptop Tablet PC Pen-and-paper
Outcomes	Acceptability: ease of use; effectiveness of the information provided by the system in helping users to complete the quality of life questionnaire; and satisfaction with the system
Notes	Questions for the usability evaluation questionnaire were extracted from IBM CSUQ

Risk of Bias

Random Sequence Generation	High risk	The procedure by which clinicians selected their patients was not specified in the study report, and almost

		half of the patients (47.30%) were allocated to the laptop group.
Allocation Concealment	Unclear risk	Not enough information available from the study report
Blinding of Participants and Personnel	High risk	The study report does not state whether participants or personnel were blinded. However, blinding might have not been possible as the type of device would immediately reveal to what group participants were allocated. The motivation to complete the self-administered survey questionnaire might have been affected if participants were aware of what other delivery modes were being offered to other participants
Blinding of Outcome Assessment	High risk	The study report does not state whether outcome assessors were blinded to the allocation of participants. However, if the calculation of final scores was necessary, there is potential for detection bias particularly for the responses collected using pen-and-paper instruments
Incomplete Outcome Assessment	Low risk	The study report suggests that all the participants that were enrolled and randomised completed the study and their data were included in the final analysis
Selective Reporting	High risk	The authors also evaluated the impact that the patient profile (both clinical and technological) had on the usability evaluation they conducted. For this, they only concentrated on the laptop group as "this is the

		only group where enough samples were collected (to assure reliable results)"
Other Bias	High risk	Although the purpose of this study was to conduct a usability evaluation of the electronic delivery of the EORTC QLQ-C30, there is a concern that the overall scores (and their SD) were not reported. In addition, the SD for the usability scores was not reported. Finally, the laptop group is over-represented in this study

Appendix 25. Characteristics of Newell and colleagues (2015)

Methods	*Study design*: RCT *Country*: US *Incentives provided*: USD 40 gift card *Type of device & platform*: iPad 2 (iOS) *Functionality*: none mentioned *Human Computer Interaction*: none mentioned *Sampling schedule*: one survey completion session *Additional interventions*: all participants received a tutorial on the operation of the tablet computer (iPad 2)
Data	*Name of survey*: Center for Epidemiological Studies Depression Scale (CES-D); Regulatory Focus Questionnaire (RFQ) *Validation status*: validated *Application of the survey*: mental health assessment; personality assessment *Population*: healthy adults *Age group*: adults, mean age 55.8 years (SD 11.9) *Gender composition of the sample*: 59% of female participants *Setting of survey completion*: clinical
Comparisons	Pen-and-paper
Outcomes	Equivalence: mean score differences Acceptability: ease of use; clarity of items; preference; perceived ability to complete a survey questionnaire; and satisfaction
Notes	Participants were drawn from two counties meeting the state of Florida's statutory definition of a rural community with 100 inhabitants or fewer per square mile The study authors oversampled participants from a black ethnic background in order to have "a representative sample of those who were documented to be disadvantaged and living in the rural South" There was a double randomisation procedure. Participants were randomised to complete the first set of questions using either an iPad or pen-and-paper. Participants were subsequently randomised to complete the second survey

		questionnaire using either an iPad or pen-and-paper. For acceptability, the comparison between delivery modes was made between those who completed both sets of survey questionnaires using an iPad (CES-D and RFQ), and those who completed both sets of questionnaires via pen-and-paper
Risk of Bias		
Random Sequence Generation	Low risk	Randomisation was conducted using Qualtrics software
Allocation Concealment	Unclear risk	The study report does not state if allocation of participants was concealed
Blinding of Participants and Personnel	High risk	The study report does not state whether participants or personnel were blinded. However, blinding might have not been possible as the delivery mode (iPad or paper) would have immediately revealed to what group participants had been allocated. In addition, all participants (regardless of their allocation) received a tutorial on how to use an iPad. The motivation to complete the self-administered survey questionnaire might have been affected due to participants' awareness of the alternative delivery mode
Blinding of Outcome Assessment	High risk	The study report does not state whether outcome assessors were blinded to the allocation of participants. However, since the calculation of final scores was necessary, there is potential for detection bias
Incomplete Outcome Assessment	Low risk	The study report suggests that all the participants that

		were enrolled and randomised completed the study and their data were included in the final analysis
Selective Reporting	High risk	For acceptability, the analyses were conducted by comparing those participants who completed both the CES-D and the RFQ on an iPad against those who complete the same survey questionnaires using pen-and-paper (thus excluding those who completed one scale on an iPad and the other one using pen-and-paper) Participants in the second community (but not the first) were asked to complete the BRIEF health literacy scale, to report on their technological experience, and to complete additional survey format items. According to the study authors the "rationale for these additions was to enable a more complete description of the sample"
Other Bias	High risk	All participants received a tutorial on how to use an iPad; this could have acted as an intervention in its own right. In addition, participants in the second community used an iPad to complete the BRIEF scale and to complete additional items

Appendix 26. Characteristics of Salaffi and colleagues (2013)

Methods	*Study design*: crossover trial *Country*: Italy *Incentives provided*: none reported *Type of device & platform*: app running on an Archos 101 (Android OS) tablet *Functionality*: all questions were compulsory; users were unable to see the next question until they had answered the current one *Human Computer Interaction*: one question per screen with both visual and auditory stimuli; data were saved automatically; voice and text synchronisation; replay buttons for the question stem and the individual response options *Sampling schedule:* one survey completion session taking place on a single day *Additional interventions*: training by research staff
Data	*Name of survey*: Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and Bath Ankylosing Spondylitis Functional Index (BASFI) *Validation status*: validated *Application of the survey*: functional status assessment *Population*: 55 patients diagnosed with axial spondyloarthritis *Age group*: adults, mean age 51 years (range 34 to 63 years) *Gender composition of the sample*: 81.82% male participants; 18.18% female participants *Setting of survey completion*: clinical
Comparisons	Pen-and-paper
Outcomes	Equivalence: mean score differences; and ICC Time to completion: mean time taken to complete the questionnaire Acceptability: acceptance; preference
Notes	Test-retest reliability was assessed in this study
Risk of Bias	

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	High risk	The presence of carry-over effect was not explored in this study; however, the washout period does not seem adequate to minimise the likelihood of carry-over effect (60 minutes). The authors stated that recall bias was reduced by organising various activities, such as visiting the physician during the interval
First period data available	Low risk	Data from both sampling periods were included in the statistical analyses
Correct Analysis	Low risk	Statistical methods are appropriate for assessing data equivalence between delivery modes. There appears to be some confusion around the reporting of SD of the mean scores.
Comparability of results with those from parallel trials	Unclear risk	Randomisation procedure followed was not reported; presence of carry-over effect was not explored; short washout period

Appendix 27. Characteristics of Schemmann and colleagues (2013)

Methods	*Study design*: crossover trial *Country*: Germany *Incentives provided*: none mentioned *Type of device & platform*: app running on tablet device *Functionality*: not specified *Human Computer Interaction*: not specified *Sampling schedule*: one survey completion session taking place on a single day *Additional interventions*: none specified
Data	*Name of survey*: International Hip Outcome - Short Version (iHOT-12) *Validation status*: validated *Application of the survey*: functional status assessment *Population*: 60 patients being treated with hip arthroscopy *Age group*: adults *Gender composition of the sample*: not specified *Setting of survey completion*: clinical
Comparisons	Pen-and-paper
Outcomes	Equivalence: ICC Acceptability: ease of use; preference
Notes	Type of publication: abstract

Risk of Bias

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Unclear risk	It is unclear if the authors tested for the presence of carry-over effect, and it is unclear how long the washout period was
First period data available	Unclear risk	It is unclear if data from both sampling periods were included in the analyses
Correct Analysis	Unclear risk	It is not possible to assess the appropriateness of the statistical analyses
Comparability of results with those from parallel trials	Unclear risk	Not enough information about the randomisation procedure, presence of carry-over effect, or length of the washout period

Appendix 28. Characteristics of Sigaud and colleagues (2014)

Methods	*Study design*: crossover trial *Country*: France *Incentives provided*: none mentioned *Type of device & platform*: app running on a smartphone device *Functionality*: not reported *Human Computer Interaction*: not reported *Sampling schedules*: 3 months (frequency not reported) *Additional interventions*: none mentioned
Data	*Name of survey*: diary for monitoring of treatment *Validation status*: non-validated *Application of the survey*: diary *Population*: 29 patients diagnosed with severe Haemophilia A and treated with recombinant Factor VIII *Age group*: adults, mean age 27.7 years *Gender composition of the sample*: 100% male participants *Setting of survey completion*: naturalistic
Comparisons	Pen-and-paper
Outcomes	Rate of diary completion (adherence to data collection protocols) Acceptability: satisfaction; and willingness to use the delivery mode
Notes	Type of publication: abstract

Risk of Bias

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Low risk	The study authors found that "the sequence of the two diary supports and the specific effect related to the patient had no significant impact on the diary completion (p=0.1960 and p=0.5552, respectively)"
First period data available	Unclear risk	Insufficient information to determine if data from both sampling periods were included in the statistical analyses

Correct Analysis	Unclear risk	Insufficient information to determine the appropriateness of the statistical analyses
Comparability of results with those from parallel trials	Unclear risk	Not enough information about the randomisation procedure used, tests for carry over effect, and the duration of the washout period

Appendix 29. Characteristics of Stomberg and colleagues (2012)

Methods	*Study design*: RCT *Country*: Sweden *Incentives provided*: none reported *Type of device & platform*: MediPal, smartphone app running on multiple devices (iOS, Android, and Java-enabled devices) *Functionality*: configurable number/times of questions displayed on each day, and configurable questions *Human Computer Interaction*: push notifications displayed every 4 hours (alerts), SMS reminders, one question per page, questions disappeared as soon as an answer was provided, and data were saved automatically *Sampling schedule:* four times a day for 6 days *Additional interventions*: training, phone calls during the data collection period allowed, and installation of the app by research staff
Data	*Name of survey*: self reported post-surgical pain *Validation status*: non-validated instrument *Application of the survey*: pain assessment *Population*: patients undergoing planned surgery (vaginal hysterectomy or laparoscopic cholecystectomy) *Age group*: adults, mean age 46.5 years old, range 18 to 66 years *Gender composition of the sample*: 87.5% female participants, 12.5% male participants *Setting of survey completion*: naturalistic
Comparisons	Pen-and-paper
Outcomes	Equivalence Adherence to the data collection protocol
Notes	Response rate was measured as compliance with the original data collection protocol

Risk of Bias

Random Sequence Generation	Unclear risk	The study report states that participants were randomly allocated to the experimental

		groups; however, the specific procedure was not mentioned
Allocation Concealment	Unclear risk	The study report does not state if allocation of participants was concealed
Blinding of Participants and Personnel	High risk	The study report does not state whether participants or personnel were blinded. However, blinding might have not been possible as the type of device would immediately reveal to what group participants were allocated. The motivation to complete the self-administered survey questionnaire might have been affected if participants were aware of what other delivery mode was being offered to other participants: "most of them were interested in the mobile phone system, and some expressed disappointment on being allocated to the control group"
Blinding of Outcome Assessment	High risk	The study report does not state whether outcome assessors were blinded to the allocation of participants. However, if the calculation of final scores was necessary, there is potential for detection bias for the pain assessments collected using the pen-and-paper instrument.
Incomplete Outcome Assessment	High risk	Three participants were not included in the final statistical analysis, and it is unclear how the study authors dealt with this. Moreover, some participants did not submit data
Selective Reporting	High risk	Response rate was measured in relation to participants' adherence to the original data collection protocol. Moreover, the unit of reporting of this outcome changed from number of responses obtained to percentage; however, it is not entirely clear what the

		latter represents. Overall reporting of pain scores was done according to the type of surgery; however, it was not reported at the experimental group level. Nevertheless, daily pain scores were reported according to group allocation. The authors only included 'correct' responses in their analysis: participants responding correctly at all of the times of the days specified in the data collection protocol
Other Bias	High risk	Participants in the intervention group received training on both pain management and use of the mobile technology. When compared to the information received by patients in the control group, this could potentially have acted as an intervention in itself, resulting in more engagement from patients in the intervention group. Furthermore, one of the purpose was the evaluation of a commercial product. Finally, the sampling period was not the same for both groups

Appendix 30. Characteristics of Sun and colleagues (2013) (A)

Methods	*Study design*: crossover trial *Country*: Canada *Incentives provided*: none mentioned *Type of device & platform*: app running on a smartphone device *Functionality*: not reported *Human Computer Interaction*: not reported *Sampling schedule*: one survey completion session taking place during a single day *Additional interventions*: none reported
Data	*Name of survey*: Faces Pain Scale Revised (FPS - R) *Validation status*: validated *Application of the survey:* pain assessment *Population*: 40 patients undergoing surgery *Age group*: children, median age 7.5 years (range 4 to 11) *Gender composition of the sample*: not specified *Setting of survey completion:* clinical
Comparisons	Pen-and-paper
Outcomes	Equivalence: mean score differences; and Pearson's correlation Acceptability: preference
Notes	Type of publication: abstract

Risk of Bias

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Unclear risk	Insufficient information to assess this domain
First period data available	Unclear risk	Insufficient information to determine if data from both sampling periods were included in the statistical analyses
Correct Analysis	Unclear risk	Insufficient information to determine the appropriateness of the statistical analyses
Comparability of results with those from parallel trials	Unclear risk	Not enough information about the randomisation procedure used, tests for carry over effect, and the duration of the washout period

Appendix 31. Characteristics of Sun and colleagues (2013) (B)

Methods	*Study design*: crossover trial *Country*: Canada *Incentives provided:* none mentioned *Type of device & platform*: app running on a smartphone device *Functionality*: not reported *Human Computer Interaction*: not reported *Sampling schedule:* one survey completion session taking place on a single day *Additional interventions*: none reported
Data	*Name of survey*: Color Analog Scale (CAS) *Validation status*: validated *Application of the survey*: pain assessment *Population*: 60 patients undergoing surgery *Age group*: children, median age 13.5 years (range 5 to 18) *Gender composition of the sample*: not specified *Setting of survey completion:* clinical
Comparisons	Plastic
Outcomes	Equivalence: mean score differences; and Pearson's correlation Acceptability: preference
Notes	Type of publication: abstract

Risk of Bias

Suitability of crossover design	Low risk	Suitable study design as recommended by Coons and colleagues (2009)
Carry-over effect	Unclear risk	Insufficient information to assess this domain
First period data available	Unclear risk	Insufficient information to determine if data from both sampling periods were included in the statistical analyses
Correct Analysis	Unclear risk	Insufficient information to determine the appropriateness of the statistical analyses
Comparability of results with those from parallel trials	Unclear risk	Not enough information about the randomisation procedure used, tests for carry over effect, and the duration of the washout period

Appendix 32. Socio-demographic survey

1. What is your age group?
    - 18 to 22 years
    - 23 to 27 years
    - 28 to 32 years
    - 33 to 37 years
    - 38 years or older
2. How would you describe your race/ethnicity?
    - White British
    - White Irish
    - Other White
    - White and Black Caribbean
    - White and Black African
    - White and Asian
    - Other Mixed
    - Indian
    - Pakistani
    - Bangladeshi
    - Other Asian
    - Black Caribbean
    - Black African
    - Other Black or Black British
    - Chinese
    - Other Ethnic Group
    - Not stated
3. What is your marital status?
    - Single
    - Married/In civil partnership
    - Divorced/Civil partnership that has been dissolved
    - Widowed

- Separated
- Not stated

4. What is your employment status?
    - Employed, full-time
    - Employed, part-time
    - Self-employed
    - Not employed, looking for work
    - Not employed, not looking for work
    - Disability/Not able to work
5. What is your highest level of education?
    - University or college degree
    - University or college qualification below degree level
    - A Levels or equivalent
    - GCSE or equivalent
    - None of these
6. Do you own a smartphone (e.g., iPhone, Samsung, Blackberry, HTC, Sony)?
    - Yes
    - No
7. Do you own a tablet computer (e.g., iPad, iPad mini, Samsung tablet)?
    - Yes
    - No
8. Is this your first pregnancy?
    - Yes
    - No
9. (If answer to previous question is *No*) How many children have you given birth to?
10. When is your current baby due?
11. Have you ever been diagnosed with depression?
    - Yes
    - No

Appendix 33. Whooley questions

1. Over the past month, have you been bothered by feeling down, depressed or hopeless?
    - Yes
    - No
2. Over the past month, have you been bothered by having little interest or pleasure in doing things?
    - Yes
    - No

Source: Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression – two questions are as good as many. *Journal of General Internal Medicine.* 1997;12:439-45.

Appendix 34. Edinburgh Postnatal Depression Scale

As you are pregnant or have recently had a baby, we would like to know how you are feeling. Please check the answer that comes closest to how you have felt IN THE PAST 7 DAYS, not just how you feel today

In the past 7 days:

1. I have been able to laugh and see the funny side of things
    - As much as I always could
    - Not quite so much now
    - Definitely not so much now
    - Not at all
2. I have looked forward with enjoyment to things
    - As much as I ever did
    - Rather less than I used to
    - Definitely less than I used to
    - Hardly at all
3. I have blamed myself unnecessarily when things went wrong
    - Yes, most of the time
    - Yes, some of the time
    - Not very often
    - No, never
4. I have been anxious or worried for no good reason
    - No, not at all
    - Hardly ever
    - Yes, sometimes
    - Yes, very often
5. I have felt scared or panicky for no very good reason
    - Yes, quite a lot
    - Yes, sometimes
    - No, not much
    - No, not at all
6. Things have been getting on top of me
    - Yes, most of the time I haven't been able to cope at all
    - Yes, sometimes I haven't been coping as well as usual
    - No, most of the time I have coped quite well
    - No, I have been coping as well as ever
7. I have been so unhappy that I have had difficulty sleeping

- Yes, most of the time
- Yes, sometimes
- Not very often
- No, not at all

8. I have felt sad or miserable
    - Yes, most of the time
    - Yes, quite often
    - Not very often
    - No, not at all
9. I have been so unhappy that I have been crying
    - Yes, most of the time
    - Yes, quite often
    - Only occasionally
    - No, never
10. The thought of harming myself has occurred to me
    - Yes, quite often
    - Sometimes
    - Hardly ever
    - Never

Sources:

- Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *British Journal of Psychiatry.* 1987; 150:782-786.
- Wisner KL, Parry BL, Piontek CM. Postpartum Depression. *New England Journal of Medicine.* 2002; 347(3): 194-199.

Appendix 35. Momentary questions

1) I feel…
    a) Great
    b) Good
    c) Ok
    d) Not good
    e) Not good at all
2) How rested do I feel?
    a) Very rested
    b) Somewhat rested
    c) Neither rested nor tired
    d) Somewhat tired
    e) Very tired
3) How worried do I feel?
    a) Very worried
    b) Quite worried
    c) Somewhat worried
    d) A little worried
    e) Not at all worried
4) Am I enjoying myself?
    a) A lot
    b) Quite a lot
    c) Somewhat
    d) Not much
    e) Not at all
5) How energetic do I feel?
    a) Very energetic
    b) Fairly energetic
    c) Moderately energetic
    d) Not very energetic
    e) Not at all energetic

Appendix 36. Contextual questions

1) Right now, I am…
    - At home
    - At work
    - At the shops
    - Outside
    - At a Café or restaurant
    - In transport
    - At a friend's or family
    - In a clinic
    - Other
2) Right now, I am…
    - Relaxing
    - Working
    - Shopping
    - Looking after others
    - Eating or drinking
    - Travelling
    - Exercising
    - Waiting
    - None

Appendix 37. Acceptance survey

Please rate your agreement with the following statements with respect to this system:

1) This app is easy to use: Strongly agree – Strongly disagree (7-point Likert Scale)
2) I learned to use this app quickly: Strongly agree – Strongly disagree (7-point Likert Scale)
3) I would recommend this app to a friend: Strongly agree – Strongly disagree (7-point Likert Scale)
4) I would repeat the experience of using this app: Strongly agree – Strongly disagree (7-point Likert Scale)
5) I found using this app an engaging experience: Strongly agree – Strongly disagree (7-point Likert Scale)
6) I found the assessments useful: Strongly agree – Strongly disagree (7-point Likert Scale)
7) I was asked to provide reports:
    - Much too infrequently
    - Too infrequently
    - The right amount
    - Too frequently
    - Much too frequently
8) The experience of using this app met my needs: Strongly agree – Strongly disagree (7-point Likert Scale)
9) What were your motivations in using this app?
10) What did you like the most about the experience of using this app?
11) What did you like least about the experience of using this app?
12) How would you improve the experience of using this app?
13) Do you have any other comments that you would like to make?

To be answered by those participants allocated to the retrospective plus momentary assessment strategy:

14) I found it useful to compare reports made right now with those made over the past 7 days: Strongly agree – Strongly disagree (7-point Likert Scale)

Appendix 38. Overview of findings with references

Chapter 2 [1] aimed to identify delivery mode effects that are the result of administering self-reported health surveys via apps. The findings were interpreted using the survey error model [2-4]. Concerning errors of measurement, survey responses collected through apps were equivalent to those collected through other delivery modes. This was the case as long as the intended applications of a survey, the intensity of its administration, and the conditions under which it was validated remained unchanged. Similar to the recommendations of Coons and colleagues [5], the included studies claimed that if no significant changes are made to a survey during its adaptation for use with electronic devices, a full validation study is not required. Concerning errors of representation, there was no data to assess if app users were representative of the general population or of patient populations. This is important given the ever present-risk of excluding vulnerable patients whenever a new technology is introduced in care pathways.

Interestingly, the included studies did not specify the changes they made to a survey, and hence offered no guidance in terms of what constitutes a significant change. Drawing from psychology and survey methodology, this systematic review highlighted the factors that could influence survey responses when using mobile apps. These factors were grouped into two categories: survey completion setting; and mediators of delivery mode effects.

Survey completion setting

This systematic review identified two survey completion settings: clinical and naturalistic. The former refers to surveys that are completed in clinical facilities, whereas the latter refers to surveys that are completed anywhere outside these facilities [1]. Data generated in clinical settings are typically a by-product of routine clinical activities, which gives practitioners more control over data quality assurance measures [1, 6, 7]. This creates a high degree of confidence in the validity and reliability of these data even when collected through apps, which facilitates their integration into clinical workflows. In naturalistic settings, apps are able to collect momentary and ecologically valid data, which could provide more comprehensive assessments of a patient's health status [6, 8]. Integrating these data into clinical workflows

however, can be challenging due to the lack of validated surveys for use in these settings, and the characteristics of survey administration.

Concerning its characteristics, survey administration in naturalistic settings typically requires repeated measures over time. In this context, patient adherence is key to ensure data completeness. User-centred design has been proposed as an effective strategy for improving adherence with digital health interventions [9, 10]. However, its implementation in healthcare is challenging due to differences in user requirements between patients and practitioners, the heterogeneity of patient populations, the dynamic nature of many health phenomena, and time constraints. Not least because it requires an iterative approach.

Mediators of delivery mode effects

Using a survey with an app involves many decisions about survey design, choice of technology, human-computer interaction factors and sampling protocols. These factors could mediate the size and direction of delivery mode effects and bias clinical decisions. For example, automated reminders, the content and timing of an invitation email, the order of response options, the use of radio buttons, the size of text boxes for open ended questions, and the number of questions on a single screen can all affect survey responses [11-16]. For this reason, there is a need to move from overall assessments of data equivalence to more detailed assessments of these mediators.

Delivery mode effects could also result from the incompatibility between survey administration and management processes, and a clinic's workflows. Research tends to prioritise internal validity over external validity, providing useful information about the impact of apps on the properties of surveys under highly controlled circumstances. However, they offer no guidance on how to implement this technology in practice, and on the impact that implementation strategies can have on survey responses. The urgent need for this type of evidence is increasingly being acknowledged [17-19].

These issues were subsequently explored by two feasibility studies conducted in a relatively highly structured clinical context: antenatal care [20, 21]. Antenatal depression was the main focus due to its burden [36] and the availability of clinically validated screening scales [20].

Data collection in clinical settings

Chapter 3 [22] explored the effect of survey layout, as one mediator of delivery mode effects, on the responses given to the Whooley questions and the EPDS on a tablet computer app in the waiting area of antenatal clinics. There was no evidence of an association between survey layout and the proportion of participants screening positive with the Whooley questions, participants' median EPDS scores, and the proportion of participants scoring one point or more on question ten of the EPDS [23].

However, there was an association between survey layout and the proportion of participants at each EPDS scoring interval [23]. This effect was driven by the proportion of participants at the highest scoring interval of this scale. This could indicate that a paging layout could lead to pregnant women with increased risk of antenatal depression not being identified by this scale, or that a scrolling layout leads to a higher number of false positives. In either case, this carries implications for practice in terms of the psychosocial impact on patients, added pressure on services, and the potential medico-legal consequences of false negative results [24, 25]. More broadly, this effect illustrates the risks of administering a validated survey through an app on the basis of overall data equivalence only.

Furthermore, the time taken to complete the surveys did not exceed patients' waiting time, there was a low breakoff rate, and only a small proportion of participants asked for assistance (mostly to clarify the meaning of questions). Taken together, these findings suggest that this approach to antenatal depression screening did not disrupt clinics' workflows. They also suggest that mobile technology in antenatal clinics might not pose a technical challenge for pregnant women, especially given the high rates of smartphone and tablet computer ownership found in this sample of participants. However, it is important to explore if this is the case across the spectrum of socio-demographic characteristics to avoid the marginalisation of vulnerable women.

Chapter 4 evaluated how the same app could be used for assessing the diagnostic properties of the Whooley questions against different EPDS scoring thresholds for pregnant women [26]. The findings showed that the discriminating power of the Whooley questions was better when 13 points was used as the scoring threshold for the EPDS. This supports the two-stage

approach to antenatal depression screening recommended by NICE, and indicates that this EPDS scoring threshold might be more appropriate for triggering subsequent mental health care pathways [27, 28]. These findings also suggest that apps can maximise the scalability and speed of data collection for diagnostic accuracy studies. However, their integration with formal diagnostic procedures needs to be explored further.

Data collection in naturalistic settings

Chapters 5 and 6 explored (i) the feasibility of using a bespoke app (i.e., BrightSelf) running on participants own smartphones for the longitudinal administration of the EPDS, and (ii) the effect of momentary questions on participant adherence with this app [29]. Momentary questions about mood, worry, energy, sleep and enjoyment were intended to help participants to self-reflect on their mood changes during pregnancy [30]. In helping participants to achieve this goal, momentary questions were thought to be able to promote effective engagement with the app, thus improving adherence [31-33].

Three operational definitions of adherence were used to explore the tension between conflicting views on patient adherence revealed in the PPI sessions that preceded this study: patients as passive entities vs. patients as informed individuals playing an active role in their healthcare. These definitions ranged from participants having to complete all the expected assessment, to participants having to complete at least one expected assessment at any point during the study.

The findings showed low overall adherence with BrightSelf when participants were expected to complete all the expected assessments, or to complete at least one assessment during each sampling period. Adherence increased considerably however, when participants were expected to complete at least one assessment at any point during the study. This suggests that patients will use a health app running on their own phones in a way that suits their needs, and not as prescribed by healthcare practitioners. One implication of this finding is that adherence is a conglomerate of multiple patterns of utilisation, rather than a summative phenomenon [34]. Indeed, multiple patterns of adherence were identified, which ranged from completing the first expected assessment only, to completing all the expected assessments.

This highlights that in order to fully support the goals of its intended users and promote effective engagement, the development of mHealth interventions should follow a truly user-centric approach. One that is true to its iterative nature, and in which the requirements from all user groups (in this case, both patients and clinicians) are weighed equally. This is relevant for heterogeneous patient populations and for the creation of systems that support the activities of clinicians in a way that is compatible with their workflows.

Chapter 6 also revealed implementation issues associated with the deployment of BrightSelf in antenatal clinics. Most participants downloaded, installed and activated BrightSelf at the clinics. Participants who did not do so came predominantly from clinics that did not offer Wi-Fi access. Perhaps they were asked to complete the activation process at home, thus reducing their willingness to complete this task. This suggests that healthcare services need to consider and provide the most basic infrastructure to deliver apps directly onto patients' devices [10]. It also indicates that clinical staff must become familiar with relevant apps in order to be able to guide patients through installation and usage. As a result of this, clinicians might need to become active players in the development of bespoke systems or to learn about the vast number of third-party apps that are available in many clinical areas. This in turn raises issues around the interpretation and reliability of patient data that have been gathered through multiple apps, and the implications for patient safety of using unregulated apps [35].

The alerting system highlighted certain challenges for the data processing requirements of clinicians and the configuration of services. Apps for gathering patient data in naturalistic settings could generate a vast amount of data points. Clinical services do not have the capacity to process these data points in time to react appropriately. This requires the implementation of algorithms that can detect declines in health, and process and present data in a way that helps practitioners to make clinical decisions [6]. However, these decisions might require a proactive approach that also has resource implications. Namely, the transition from a paternalistic model of care to a more collaborative approach where patients are empowered to take more control of their own health [6].

Furthermore, the responses to the post-study acceptance survey suggest that patients who completed the study found BrightSelf easy to use, an enjoyable experience that met their needs, useful, and with just the right amount of assessments. However, these responses were

obtained from a small proportion of participants. This further highlights the need for more comprehensive assessments of acceptability to clearly identify which patients find this approach most beneficial.

Overall, this doctoral work demonstrated that it is feasible to use apps running on consumer mobile devices for the screening and monitoring of depression in both clinical and naturalistic settings. This work also demonstrated that adapting a survey for use via an app and incorporating mobile technology into care pathways are complex process that require evidence at multiple levels.

References

1. Marcano Belisario JS, Jamsek J, Huckvale K, O'Donoghue J, Morrison CP, Car J. Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. *Cochrane Database of Systematic Reviews*. 2015; 7:MR000042. Available from: doi: 10.1002/14651858.MR000042.pub2 [Accessed 10 October 2018].
2. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey Methodology*. 2nd Edition. New Jersey, US: Wiley; 2009.
3. Lavrakas, PJ (ed.). *Encyclopedia of Survey Research Methods*. London, UK: SAGE Publications; 2008.
4. Groves RM, Lyberg L. Total Survey Error: Past, Present, Future. *Public Opinion Quarterly*. 2010;74(5):849-879. Available from: doi: 10.1093/poq/nfq065 [Accessed 31 July 2018].
5. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, Lenderking WR, Cella D, Basch E, ISPOR ePRO Task Force. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. *Value in Health*. 2009;12(4): 419-29. Available from: doi: 10.1111/j.1524-4733.2008.00470.x [Accessed 31 July 2018].
6. West P, Van Kleek M, Giordano R, Weal MJ, Shadbolt N. Common Barriers to the Use of Patient-Generated Data Across Clinical Settings. *Proceedins of the 2018 Conference on Human Factors in Computing Systems (CHI),* 21 – 26 April 2018,

Motreal, Canada. Available from: doi: 10.1145/3173574.3174058 [Accessed 31 July 2018].

7. West P, Giordano R, Van Kleek M, Shadbot N. The Quantified Patient in the Doctor's Office: Challenges and Opportunities. *Proceedings of the 2016 Conference on Human Factors in Computing Systems (CHI),* 7 – 12 May, San Jose, California, United States. Available from: doi: 10.1145/2858036.2858445 [Accessed 31 July 2018].
8. Neff G, Nafus D. *Self-Tracking*. London, UK: The MIT Press; 2016.
9. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, Merchant GC, Naughton F, Blandford A. Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *American Journal of Preventive Medicine*. 2016;51(5):833-842. Available from: doi: 10.1016/j.amepre.2016.06.015 [Accessed 31 July 2018].
10. Lupton D. Digital health now and in the future: Findings from a participatory design stakeholder workshop. *Digital Health*. 2017;3: 1 – 17. Available from: doi: 10.1177/2055207617740018 [Accessed 22 October 2018].
11. Fan W, Yan Z. Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behaviour*. 2010;26:132-139. Available from: doi: 10.1016/j.chb.2009.10.015 [Accessed 31 July 2018].
12. Mavletova A, Couper MC. Mobile Web Survey Design: Scrolling versus Paging, SMS versus E-Mail Invitations. *Journal of Survey Statistics and Methodology*. 2014;2(4):498-518. Available from: doi: 10.1093/jssam/smu015 [Accessed 31 July 2018].
13. Conrad FG, Schober MF, Jans M, Orlowski RA, Nielsen D, Levenstein R. Comprehension and engagement in survey interviews with virtual agents. *Frontiers in Psychology*. 2015;6:1578. Available from: doi: 10.3389/fpsyg.2015.01578 [Accessed 31 July 2018].
14. Mavletova A, Couper MC. Grouping of Items in Mobile Web Questionnaires. *Fields Methods*. 2015;28(2):170-193. Available from: doi: 10.1177/1525822X15595151 [Accessed 31 July 2018].
15. Maloshonok N, Terentev E. The impact of visual design and response formats on data quality in a web survey of MOOC students. *Computers in Human Behavior*.

2016;62:506-515. Available from: doi: 10.1016/j.chb.2016.04.025 [Accessed 31 July 2018].

16. Petrovčič A, Petrič G, Manfreda KL. The effect of email invitation elements on response rate in a web survey within an online community. *Computers in Human Behavior*. 2016;56:320-329. Available from: doi: 10.1016/j.chb.2015.11.025 [Accessed 31 July 2018].
17. Devlin AM, McGee-Lennon M, O'Donnell CA, Bouamrane MM, Agbakoba R, O'Connor S, Grieve E, Finch T, Wyke S, Watson N, Browne S, Mair FS. Delivering digital health and well-being at scale: lessons during learned during the implementation of the dallas program in the United Kingdom. *Journal of the American Medical Informatics Association.* 2016;23(1):48-59. Available from: doi: 10.1093/jamia/ocv097 [Accessed 31 July 2018].
18. Agbakoba R, McGee-Lennon M, Bouamrane MM, Watson M, Mair FS. Implementation factors affecting the large-scale deployment of digital health and well-being technologies: A qualitative study of the initial phases of the 'Living-It-Up' program. *Health Informatics Journal*. 2016;22(4):867-877. Available from: doi: 10.1177/1460458215594651 [Accessed 31 July 2018].
19. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-Health: a systematic review of systematic reviews (an update). *Implementation Science*. 2016;11:146. Available from: doi: 10.1186/s13012-016-0510-7 [Accessed 31 July 2018].
20. National Collaborating Centre for Mental Health (2014 – 2015 Update). *Antenatal and postnatal mental health: clinical management and service guidance*. Clinical Guideline CG192.
21. National Institute for Health and Clinical Excellence (2008 – 2017 Update). *Antenatal care for uncomplicated pregnancies*. Clinical Guideline CG62.
22. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Morrison C, Car J. Tablet computers for implementing NICE antenatal mental health guidelines: protocol of a feasibility study. *BMJ Open*. 2016;6:e009930. Available from: doi: 10.1136/bmjopen-2015-009930 [Accessed 31 July 2018].
23. Marcano-Belisario JS, Gupta AK, O'Donoghue J, Ramchandani P, Morrison C, Car J. Implementation of depression screening in antenatal clinics through tablet

computers: results of a feasibility study. *BMC Medical Informatics and Decision Making*. 2017;17:59. Available from: doi: 10.1186/s12911-017-0459-8 [Accessed 31 July 2018].

24. Miles A, Rodrigues V, Sevdalis N. The effect of information about false negative and false positive rates on people's attitudes towards colorectal cancer screening using faecal occult blood testing (FOBt). *Patient Education and Counseling*. 2013;93(2):342-349. Available from: doi: 10.1016/j.pec.2013.06.010 [Accessed 31 July 2018].
25. Brodersen J, Siersma VD. Long-Term Psychosocial Consequences of False-Positive Screening Mammography. *Annals of Family Medicine*. 2013;11(2):106-115. Available from: doi: 10.1370/afm.1466 [Accessed 31 July 2018].
26. Littlewood E, Ali S, Ansell P, Dyson L, Gacoyne S, Hewitt C, et al. Identification of depression in women during pregnancy and the early postnatal period using the Whooley questions and the Edinburgh Postnatal Depression Scale: protocol for the Born and Bred in Yorkshire: PeriNatal Depression Diagnostic Acuracy (BaBY PaNDA) study. *BMJ Open*. 2016;6:e011223. Available from: doi: 10.1136/bmjopen-2016-011223 [Accessed 01 August 2018].
27. Gibson J, McKenzie-McHarg K, Shakespeare J, Price J, Gray R. A systematic review of studies validating the Edinburgh Postnatal Depression Scale in antepartum and postpartum women. *Acta Psychiatrica Scandinavica*. 2009;199(5):350-364. Available from: doi: 10.1111/j.1600-0447.2009.01363.x [Accessed 01 August 2018].
28. Littlewood E, Ali S, Dyson L, Keding A, Ansell P, Bailey D, Bates D, Baxter C, Beresford-Dent J, Clarke A, Gascoyne S, Gray C, Hackney L, Hewitt C, Hutchinson D, Jefferson L, Mann R, Marshall D, McMillan D, North A, Nutbrown S, Peckham E, Pervin J, Richardson Z, Swan K, Taylor H, Waterhouse B, Wills L, Woodhouse R, Gilbody S. Identifying perinatal depression with case-finding instruments: a mixed-methods study (BaBY PaNDA – Born and Bred in Yorkshire PeriNatal Depression Diagnostic Accuracy). *Health Services and Delivery Research*. 2018;6(6). Available from: doi: 10.3310/hsdr06060 [Accessed 01 January 2019].
29. Marcano-Belisario JS, Doherty K, O'Donoghue J, Ramchandani P, Majeed A, Doherty G, Morrison C, Car J. A bespoke mobile application for the longitudinal

assessment of depression and mood during pregnancy: protocol of a feasibility study. *BMJ Open*. 2017;7:e014469. Available from: doi: 10.1136/bmjopen-2016-014469 [Accessed 01 August 2018].

30. Doherty K, Barry M, Marcano-Belisario JS, Arnaud B, Morrison C, Car J, Doherty G. A Mobile App for the Self-Report of Psychological Well-Being During Pregnancy (BrightSelf): Qualitative Design Study. *JMIR Mental Health*. 2018;5(4): e10007. Available from: doi: 10.2196/10007 [Accessed 30 December 2018].
31. Torous J, Nicholas J, Larsen ME, Firth J, Christensen. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence Based Mental Health*. 2018;21(3):116-119. Available from: doi: 10.1136/eb-2018-102891 [Accessed 01 August 2018].
32. Mulvaney SA, Vaala S, Hood KK, Lybarger C, Carroll R, Williams L, et al. Mobile Momentary Assessment and Biobehavioral Feedback for Adolescents with Type 1 Diabetes: Feasibility and Engagament Patterns. *Diabetes Technology and Therapeutics*. 2018;20(7):465-474. Available from: doi: 10.1089/dia.2018.0064 [Accessed 01 August 2018].
33. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*. 2016;9:211-217. Available from: doi: 10.2147/JMDH.S104807 [Accessed 01 August 2018].
34. Morrison C, Doherty G. Analyzing Engagement with a Web-Based Intervention Platform Through Visualizing Log-Data. *Journal of Medical Internet Research*. 2014;16(11):e252. Available from: doi: 10.2196/jmir.3575 [Accessed 01 August 2018].
35. Ashall-Payne L, Leigh S. *Health Apps, the Good, the Bad, and the Ugly: The Importance of Information*. [Online] NIHR Blog. Available from: https://www.nihr.ac.uk/blogs/health-apps-the-good-the-bad-and-the-ugly-the-importance-of-information/4702 [Accessed 01 August 2018].