Population genetics of *Streptococcus pneumoniae*: the response to antibiotic and vaccine pressure

Sarah Raphaelle Deeny

*Imperial College London, Department of Infectious Disease Epidemiology*

2011

*Thesis submitted for the degree of Doctor of Philosophy of Imperial College*
Statement of Originality

I declare that the work presented in this thesis is my own and that I have acknowledged the work of others where relevant.

Acknowledgements

I have been very lucky to have been mentored by three exceptional scientists during my PhD. I feel very privileged to have worked with such good humoured, personable and inspirational researchers, from whom I have learnt an enormous amount. Firstly I would like to thank my supervisor Christophe Fraser for his advice, assistance, encouragement and insight. I am also very grateful to my supervisor Brian Spratt whose guidance and experience was an invaluable asset throughout my PhD and particularly in preparing this thesis. Great thanks are also due to Bill Hanage, who generously provided essential comments and advice on all the research undertaken as part of this thesis.

For Chapter 2, thanks go to Marcus Shepherd and Felix Greaves for their discussions and insight. I am very grateful to Marc Lipsitch, whose earlier work and hypothesis formed the starting point for the research described in Chapters 3 and 4 and whose comments and advice on this work was gratefully received. In addition Chapters 3 and 4 benefitted from comments from Caroline Colijn and Ted Cohen and the research could not have been undertaken if data had not been generously shared by Osman Abdullahi and colleagues in Kenya and Matthew Moore and colleagues in the CDC, United States. Data analysed in Chapter 7 were generously provided by Bill Hanage.

I’d like to thank other teachers who I’ve been lucky enough to have met during my education to date, Sister Hildegard and Mrs. Porter who gave me a love of learning and science. I’m really appreciative of all the wonderful friends I made during my time in Imperial, particularly those whom I first met during my MSc and formed a ready mutual support group throughout the PhD. Huge thanks go to all my family for supporting me throughout my education. Particular thanks to my aunts Peggy, Geraldine and the McKeown sisters who were always very supportive during my move to London and throughout my PhD. Great thanks go to my brother Colm and sister Gabrielle who always helped me keep things in perspective. I am forever in the debt of my wonderful parents who have always been a source of inspiration and supported me in everything I have wanted to do and without whom this PhD could not have been completed. Lastly I’d like to thank Cian, who more than most shared in the ups and downs of this research and always encouraged me every step of the way.

This research was funded by the Medical Research Council and Imperial College London Faculty of Medicine.
Abstract

Streptococcus pneumoniae (pneumococcus) is a pathogen and a commensal of the upper respiratory tract, which is a leading cause of child mortality. The development of a successful vaccine against pneumococcal carriage and disease and the ongoing challenge of antimicrobial resistant strains mean that there is an imperative to understand how the pneumococcal population responds to vaccine and antibiotic pressure.

Strains of pneumococci with reduced susceptibility to penicillin and other antibiotics have emerged and are a cause of concern, though their clinical impact is unclear. I perform a meta-analysis to examine the impact of antibiotic non-susceptibility on the risk of mortality and show that a meningitis infection with penicillin non-susceptible pneumococci increases the risk of mortality two fold.

I then examine why some serotypes of pneumococcus are more likely to carry resistance than others. Using a mathematical model I generate the hypothesis that serotypes with a long duration of carriage are more likely to have a high prevalence of antibiotic resistance than those with a short duration of carriage. Using maximum likelihood estimation, I show that in children less than two years of age, penicillin resistance only occurs in those serotypes whose duration of carriage is nineteen days or more.

Having considered the circumstances under which antibiotic resistance carries a selective advantage in a pneumococcal serotype, I then consider the effect that the spread of an advantageous gene has on the genetic diversity in a generalised bacterial population. Most simple models predict that when a novel allele arises in a bacterial population that is fitter than other alleles at that locus, it and its descendents will increase in frequency and sweep to fixation in the population. In the absence of recombination, all genetic diversity at all loci other than those within the sweeping genotype is lost. I consider whether asymmetric recombination can prevent the loss of diversity over the whole genome. I show that asymmetric recombination, when occurring at rates estimated to date from bacterial populations, is not frequent enough to prevent the extinction of alleles from the wild-type population. Finally, I analyse sequence data sampled from carried pneumococci to examine the impact of vaccination on genetic diversity in a pneumococcal population, an example of a selective event in a pneumococcal population.
Table of Contents

STATEMENT OF ORIGINALITY ........................................................................................................ 2

ACKNOWLEDGEMENTS...................................................................................................................... 2

TABLE OF CONTENTS ......................................................................................................................... 4

LIST OF FIGURES .................................................................................................................................. 7

LIST OF TABLES .................................................................................................................................... 9

LIST OF ABBREVIATIONS ...................................................................................................................... 10

1 INTRODUCTION ............................................................................................................................... 12

1.1 Biology ........................................................................................................................................ 12

1.2 Interventions ................................................................................................................................. 18

1.3 Typing pneumococci .................................................................................................................... 24

1.4 Aims of thesis ............................................................................................................................... 27

2 THE IMPACT OF PNEUMOCOCCAL ANTIBIOTIC RESISTANCE ON MORTALITY: A REVIEW AND META-ANALYSIS .................................................................................................................. 29

2.1 Introduction .................................................................................................................................. 29

2.2 Method ......................................................................................................................................... 32

2.3 Results .......................................................................................................................................... 35

2.4 Discussion .................................................................................................................................... 48

2.5 Conclusion .................................................................................................................................... 52

3 EXPLAINING HETEROGENEOUS PREVALENCE OF RESISTANCE AMONG SEROTYPES: HYPOTHESIS AND MATHEMATICAL MODEL .............................................................................................................. 53

3.1 Introduction .................................................................................................................................. 53

3.2 Hypothesis - distribution of antibiotic resistance amongst serotypes ........................................ 56

3.3 Mathematical model of antibiotic resistance .................................................................................. 59
3.4 Results ........................................................................................................................................... 68
3.5 Exploration of key model assumptions and sensitivity analysis ..................................................... 76
3.6 Conclusions .................................................................................................................................... 80

4 ANALYSIS OF PNEUMOCOCCAL POPULATIONS FROM INVASIVE DISEASE AND CARRIAGE .............................................................................................................................. 81
4.1 Introduction ..................................................................................................................................... 81
4.2 Data sets ......................................................................................................................................... 81
4.3 Description of statistical models and statistical techniques ............................................................... 83
4.4 Results ............................................................................................................................................ 85
4.5 Discussion and analysis limitations ................................................................................................ 96
4.6 Discussion and conclusion ............................................................................................................... 100
4.7 Conclusion ..................................................................................................................................... 102

5 SELECTION AND RECOMBINATION IN BACTERIAL POPULATIONS: LITERATURE REVIEW ................................................................................................................................... 103
5.1 Introduction ..................................................................................................................................... 103
5.2 Review of selection in bacterial populations .................................................................................... 105
5.3 Two-Locus Model of a selective sweep with asymmetric recombination ....................................... 116
5.4 Conclusion ..................................................................................................................................... 122

6 CAN BACTERIAL RECOMBINATION DISRUPT THE HITCH-HIKING EFFECT OF A SELECTIVE SWEEP? .................................................................................................................. 125
6.1 Introduction ..................................................................................................................................... 125
6.2 Stochastic Two Locus Model ........................................................................................................... 131
6.3 Eight Locus Model of a Selective Sweep in a Mutating and Recombining Population ..................... 143
6.4 Clonal interference .......................................................................................................................... 153
6.5 Discussion and conclusion ............................................................................................................... 168

7 IMPACT OF PNEUMOCOCCAL CARRIAGE VACCINATION ON THE POPULATION STRUCTURE OF STREPTOCOCCUS PNEUMONIAE ........................................................................... 173
List of Figures

FIGURE 1.1 SCHEMATIC OF THE PNEUMOCOCCAL GENOME ......................................................... 14
FIGURE 2.1 FOREST PLOT OF POOLED ORS GROUPED BY SYNDROME ............................................. 44
FIGURE 2.2 GEOGRAPHICAL GROUPING OF ORS ............................................................................. 46
FIGURE 3.1 RELATIVE IMPACT OF RESISTANCE. .............................................................................. 57
FIGURE 3.2 SCHEMATIC ILLUSTRATION OF TWO THEORETICAL MODELS .................................. 60
FIGURE 3.3 PROPORTION RESISTANT; EFFECT OF TREATMENT ..................................................... 69
FIGURE 3.4 COST OF RESISTANCE: RATES OF ANTIBIOTIC CONSUMPTION .................................. 70
FIGURE 3.5 COST OF RESISTANCE: DURATION OF CARRIAGE ...................................................... 70
FIGURE 3.6 MODEL B, PROPORTION RESISTANT; EFFECT OF TREATMENT ............................... 73
FIGURE 3.7 MODEL B, COST OF RESISTANCE: EFFECT OF TREATMENT .................................. 74
FIGURE 3.8 MODEL B, COST OF RESISTANCE: DURATION OF CARRIAGE .................................. 74
FIGURE 3.9 p-sensitivity analysis ........................................................................................................ 78
FIGURE 4.1 DURATION OF CARRIAGE PER SEROTYPE ................................................................. 83
FIGURE 4.2 PREVALENCE OF PENICILLIN RESISTANCE PRE 2000 IN INVASIVE DISEASE ISOLATES FROM INDIVIDUALS (A) UNDER TWO YEARS OF AGE AND (B) OVER TWO YEARS OF AGE .......................................................... 88
FIGURE 4.3 PREVALENCE OF ERYTHROMYCIN RESISTANCE PRE 2000 IN INVASIVE DISEASE ISOLATES FROM INDIVIDUALS (A) UNDER TWO YEARS OF AGE AND (B) OVER TWO YEARS OF AGE .......................................................... 89
FIGURE 4.4 PREVALENCE OF PENICILLIN RESISTANCE. MEAN PROPORTION OF SEROTYPES PENICILLIN RESISTANT WITH 95% CONFIDENCE BOUNDS, PLOTTED AGAINST THE DURATION OF CARRIAGE. A) SEROTYPES FROM SURVEILLANCE OF INVASIVE DISEASE IN INDIVIDUALS UNDER 2 YEARS OF AGE COLLECTED BEFORE 2000, B) SEROTYPES FROM SURVEILLANCE OF INVASIVE DISEASE IN INDIVIDUALS OVER 2 YEARS OF AGE COLLECTED BEFORE 2000. THE SOLID LINE PLOTS THE CURVE FROM SIGMOID FITTED TO THE PREVALENCE OF PENICILLIN RESISTANCE (MIC ≥ 2), THE DASHED LINE, THE NULL HYPOTHESIS ........................................... 92
FIGURE 4.5 PREVALENCE OF ERYTHROMYCIN RESISTANCE ................................................................ 93
FIGURE 5.1 SCHEMATIC DIAGRAM OF BATCH CULTURE TECHNIQUE .............................................. 106
FIGURE 5.2 SCHEMATIC OF A CHEMOSTAT .................................................................................... 110
FIGURE 5.3 TWO LOCUS MODEL OF SELECTIVE SWEEP ................................................................ 118
FIGURE 5.4 COMPARISON BETWEEN SYMMETRIC AND ASYMMETRIC MODELS ...................... 121
FIGURE 6.1 STABLE ECOTYPE MODEL. ECOTYPE FORMATION EVENTS ARE REPRESENTED BY A CHANGE IN COLOUR ................................................................. 127
FIGURE 6.2 TWO LOCUS ASYMMETRIC DETERMINISTIC MODEL ............................................... 131
FIGURE 6.3 DISTRIBUTION OF FITNESS EFFECTS ............................................................................ 137
FIGURE 6.4 RELATIONSHIP BETWEEN RECOMBINATION RATE, SELECTION COEFFICIENT AND PROPORTION OF MX IN THE FINAL POPULATION IN THE TWO LOCUS STOCHASTIC MODEL .................................................. 138
FIGURE 6.5 IMPACT OF RECOMBINATION OVER TIME ................................................................. 141
FIGURE 6.6 RELATIONSHIP BETWEEN THE FIXATION PROBABILITY OF THE M ALLELE AND P(MX), THE PROPORTION OF MX IN THE POST-SWEEP POPULATION (Y AXIS) .................................................................................. 142
FIGURE 6.7 EIGHT LOCUS INDIVIDUAL MODEL OF SELECTION. THIS ILLUSTRATION SHOWS THE MODEL FOR A POPULATION OF FIVE INDIVIDUALS ........................................................................................................... 145
FIGURE 6.8 INCREASING RECOMBINATION IN A POPULATION WITH CONSTANT HOMOZYGOSITY ................................................................................................................................. 146
FIGURE 6.9 EFFECT OF MUTATION AND RECOMBINATION ON STRAIN DIVERSITY ..................... 147
List of Tables

TABLE 2.1 SEARCH TERMS .................................................................................................................. 33
TABLE 2.2 EXCLUDED STUDIES ....................................................................................................... 37
TABLE 2.3 STUDIES INCLUDED IN META-ANALYSIS ......................................................................... 38
TABLE 2.4 ADJUSTED OR ...................................................................................................................... 47
TABLE 3.1 COST OF RESISTANCE RELATIONSHIP BETWEEN PARAMETERS IN MODEL A AND MODEL B ........................................................................................................... 65
TABLE 3.2 PARAMETER TABLE FOR MODEL A AND MODEL B ............................................................... 67
TABLE 3.3 Dₜ (LOWEST DURATION OF CARRIAGE IN DAYS WITH 5% OF STRAINS RESISTANT TO ANTIBIOTICS) .................................................................................................................. 71
TABLE 3.4 Dₜ IN DAYS FOR CHANGING VALUES OF K₀ ........................................................................... 77
TABLE 3.5 Dₜ FOR CHANGING VALUES OF RELATIVE TO .................................................................... 78
TABLE 4.1 FORMER AND CURRENT CLINICAL AND LABORATORY STANDARDS INSTITUTE PENICILLIN AND ERYTHROMYCIN MINIMUM INHIBITORY CONCENTRATION (MIC) BREAKPOINTS .................................................................................. 82
TABLE 4.2 FREQUENCY AND CARRIAGE DURATION OF SEROTYPES IN PATIENTS UNDER AND OVER TWO YEARS OF AGE ................................................................. 87
TABLE 4.3 MAXIMUM LIKELIHOOD ANALYSIS: ASSOCIATION BETWEEN PREVALENCE OF PENICILLIN RESISTANCE (MIC ≥2) AND SEROTYPE DURATION OF CARRIAGE .................................................................................. 94
TABLE 4.4 MAXIMUM LIKELIHOOD ANALYSIS: ASSOCIATION BETWEEN PREVALENCE OF ERYTHROMYCIN RESISTANCE (MIC ≥1) AND SEROTYPE DURATION OF CARRIAGE .................................................................................. 95
TABLE 6.1 PROBABILITY OF FIXATION OF NEUTRAL ALLELE IN 1000 SELECTIVE SWEEPS AS A FUNCTION OF RECOMBINATION AND SELECTION ........................................................................ 139
TABLE 7.1 ORIGINAL DATA-SET COMPOSITION OF UN-VACCINATED COMMUNITIES ......................... 178
TABLE 7.2 SAMPLE CHARACTERISTICS OF VACCINATED AND UNVACCINATED COMMUNITIES ............................................................ 181
TABLE 7.3 FREQUENCY OF STS ASSOCIATED WITH VT AND NVT SEROTYPES IN 2001 AND CHANGE IN FREQUENCY IN LATER YEARS .................................................. 190
TABLE 7.4 CAPSULAR SWITCHING ........................................................................................................ 190
List of Abbreviations

**ABCS** Active Bacterial Core Surveillance

**AOM** Acute Otitis Media

**CDC** Centers for Disease Control and Prevention

**CI** Confidence Interval

**CLSI** Clinical and Laboratory Standards Institute

**CPS** Capsular Polysaccharide

**CSF** Cerebrospinal Fluid

**DNA** Deoxyribonucleic acid

**EARSS** European Antimicrobial Resistance Surveillance System

**Epi-DSS** Kilifi Epidemiological and Demographic Surveillance Study

**ESAC** European Surveillance of Antimicrobial Consumption

**GT** Glycosyltransferase

**IAM** Infinitely Many Alleles Model

**IPD** Invasive Pneumococcal Disease

**IPnI** Invasive Pneumococcal Infection

**LD** Linkage Disequilibrium

**MIC** Minimum Inhibitory Concentration

**MLST** Multi Locus Sequence Typing

**MRCA** Most Recent Common Ancestor

**NVT** NonVaccine Serotypes

**OR** Odds Ratio

**PBP** Penicillin Binding Protein

**PCR** Polymerase Chain Reaction

**PCV7** Seven valent conjugate polysaccharide vaccine

**PNSP** Penicillin NonSusceptible Pneumococci

**PRP** Penicillin Resistant Pneumococci
PSP Penicillin Susceptible Pneumococci
ST Sequence Type
VT Vaccine Serotypes
WHO World Health Organisation
1 Introduction

*Streptococcus pneumoniae* is a normal resident of the human upper respiratory tract where it is carried asymptomatically. It is an occasional but important human pathogen which can cause a wide range of disease (1). *Streptococcus pneumoniae*, also referred to as the pneumococcus, is a causative agent of diseases such as acute otitis media, meningitis and pneumonia (2). It has been estimated that 11% of all deaths of children under five in the year 2000 were caused by pneumococcal disease (3).

The development of a successful vaccine against pneumococcal carriage and disease and the ongoing challenge of antimicrobial resistant strains mean that there is an imperative to understand how public health interventions affect the evolution of the pneumococcal population. The aim of this thesis is to examine the impact of antibiotic resistance and vaccination on pneumococcal population genetics. In this chapter I review relevant aspects of the biology, epidemiology and clinical impact of the pneumococcus.

1.1 Biology

1.1.1 Capsular Polysaccharide

The pneumococcus produces an outer polysaccharide capsule, which is exposed on the surface of the pneumococcal cell, and provides a target for host immune response. Currently over 90 serotypes have been described, each with an immunochemically distinct capsular polysaccharide (CPS). Each CPS is composed of a series of repeated units of simple sugar structures (monosaccharide) linked together by glycosidic bonds, the repeated units being linked together and attached to the cell wall of the pneumococcus (4,5). Antigenic diversity between serotypes is due to the number and nature of monosaccharide structures, the type and configuration of the linkages between monosaccharides, non-sugar composition and features such as acetylation (the addition of an acetyl group to a monosaccharide), and the form of linkages between repeated units. The CPS of some serotypes are relatively simple in structure, such as serotypes 3 and 37, whereas others are more complex, such as serotypes 6A and 15C (4).

The pneumococcal capsular serotype is synthesised by the Wzx/Wzy-dependent pathway (with the exception of serotypes 3 and 37 which are synthesised by the synthase pathway). The Wzx/Wzy-dependent pathway involves the synthesis of the repeated units of the CPS, its translocation across the cytoplasmic membrane, and its polymerisation and linkage to the cell wall (6). As illustrated in Figure 1.1, the genes required for the synthesis of CPS through the Wzx/Wzy-dependent pathway
are located in a cluster of genes called the capsular biosynthetic locus (cps). The cps locus is at the same position on the pneumococcal chromosome in each serotype, bounded by two genes (dexB and aliA) which play no part in the synthesis of the capsule (4). Published sequences of cps loci have shown that although CPS are structurally diverse, the cps locus has a number of common features shared among serotypes. The first four genes of the locus (designated as wzg, wzh, wzd and wze using the Bacterial Polysaccharide Gene Naming (BPGN) scheme) are highly conserved between serotypes and are present at the 5’ end of the cps locus. In most serotypes the 5th gene encodes the initial sugar phosphate transferase WchA (also known as cpsE), which is responsible for the linkage of an activated glucose phosphate to the lipid carrier. The next section of the locus, downstream from the 5th gene, is a serotype specific region. In each serotype this region contains glycosyltransferase (GT) genes, the CPS repeat unit polymerase gene (wzw) and the CPS repeat unit flippase gene (wzx) and genes that encode acetyl transferases. Genes that code for sugar pathway proteins are often located at the 3’ end of the cps locus (7,8).
1.1.2 Recombination

Pneumococci are naturally competent bacteria, meaning that they can actively take free (extracellular) DNA from their environment, transfer it into the cell and incorporate it into their genome by homologous recombination (9). The term ‘natural transformation’ has been coined to describe this process. Although detected over fifty years ago in laboratory experiments, the important role it plays in natural populations of pneumococci (those colonising the nasopharynx of humans) only came to light in the 1990s (10).

Broadly, natural transformation occurs in three basic steps: association of free DNA with the competent pneumococcal cell, translocation of the DNA into the cytoplasm, and incorporation of the donor DNA into the genome. Free double stranded DNA associates with competent cells to form a complex that is resistant to gentle washing though not to degradation by enzymes that digest DNA (DNase) (9). Pneumococci can potentially take up any free DNA (11), although the probability of...
integration of a DNA strand decreases as genetic divergence between the donor and recipient increases (10).

Recombination plays an important role in generating genetic diversity in the pneumococcal population. Unlike mutation where changes in the nucleotide code may result in the novel allele being non-functional, homologous recombination is expected to introduce functional sequences into the recipient pneumococcal genome as the donor DNA comes from a functional pneumococcus. In addition several independent fragments (1-5% of the genome) can be taken up per pneumococcal cell allowing novel combinations of genes to arise and the simultaneous substitution of components of protein complexes can occur in one transformation event (11). Homologous recombination allows pneumococci to transfer DNA from pneumococci and closely related oral streptococci that share the same ecological niche (the human nasopharynx) (12). Thus it provides an extremely important mechanism allowing the rapid evolution of the pneumococcal genome (10). Two well studied and clinically important aspects of pneumococcal evolution where homologous recombination has played an integral role are: firstly, the creation and spread of antibiotic resistant genes; and secondly, capsular polysaccharide switching between strains, which will be examined in more detail in Section 1.2.

1.1.1 Carriage and disease

The upper respiratory tract is the ecological niche for many bacterial species which colonise the nasopharynx within the first months of life. A broad variety of organisms, including the pneumococcus, Moraxella cattarrhalis, Haemophilus influenzae, Neisseria meningitidis and Staphylococcus aureus will colonise an individual in their lifetime, commonly resulting in asymptomatic carriage, or more rarely disease (13).

1. Biology of carriage

Adhesion is the initial event in asymptomatic carriage, and requires adherence of a pneumococcal cell to the epithelial lining of the respiratory tract (14). Adhesion of the pneumococcus to the epithelial lining is mediated by pneumococcal cell wall associated surface proteins which bind to the cell surface carbohydrates of the host (15). Pneumococci have a phase shift which allows them to exist in a transparent and an opaque state; pneumococci in transparent phase state have thick cell walls but sparse capsules, while in the opaque state have thin walls but thicker capsules. Animal models of nasopharyngeal carriage have demonstrated that acquisition of nasopharyngeal carriage is achieved more readily with transparent phase state pneumococci (16). As the transparent state
has thick cell walls but a sparse capsule, it is thought that this allows the sub-capsular proteins on the cell wall to attach to the epithelial lining more efficiently (17).

The duration of asymptomatic carriage in an individual is age and serotype dependent, and although not fully understood, may be influenced by interactions between pneumococci and host immune effectors (18). Immune mediated clearance from the nasopharynx involves both antibody dependent and independent mechanisms of immunity. The antibody independent mechanisms are thought to involve a T cell response which results in the recruitment of neutrophils to the site of carriage and the subsequent clearance of asymptomatic carriage (19,20). Although the role of the polysaccharide capsule in carriage is unclear, heavily encapsulated serotypes can avoid being destroyed by phagocytes (18).

During the first two years of a child’s life there may be multiple instances of pneumococcal carriage with the same serotype or multiple different serotypes (21). In animal models of carriage there is no clear evidence that carriage of one serotype prevents subsequent carriage by the same serotype (22), although some mouse models of carriage show some evidence of immunity to re-infection (23). Thus, although it is observed that the incidence of pneumococcal carriage declines with age, the role of specific and natural human immune response (in particular in the context of exposure to multiple serotypes) is unclear (24). However as I will later describe, there is good evidence that immunity artificially induced by serotype specific vaccines is protective against carriage of certain serotypes.

ii. Epidemiology of carriage

Person to person spread of pneumococcal carriage is thought to be through direct contact between individuals (25). Transmission and carriage prevalence are highest in young children (24). Although carriage prevalence varies between studies there is a common pattern: infants can be colonised in the first few weeks after birth, and the prevalence of carriage rises during the first year of life (1,14). In most studies prevalence of carriage declines after the age of four and decreases further into adulthood (26).

There is some evidence that in healthy populations other risk factors additional to age determine the frequency of carriage. Independent determinants for nasopharyngeal carriage include ethnicity, crowding and socioeconomic factors (24). In young children, day care visits are significantly associated with carriage, and there is evidence of genotypic clustering of isolates among day care attendees, suggesting that transmission occurs among day care attendees (14,27). In a study of carriage in children under the age of three, children who attended day care had a mean prevalence
of carriage of 58%, while children who did not attend day care had a mean prevalence of carriage of 37% (27).

Studies from low income settings have shown that prevalence is higher in both children and adults in these settings compared to high income settings (21). In a study carried out in the Gambia, researchers followed a cohort of infants from birth to 11 months. During this time all the infants carried pneumococci in the nasopharynx at least once, the mean prevalence of carriage was 85% in children under 11 months (21). In contrast, in a longitudinal household study carried out in the United Kingdom which followed children and their households for 10 months, only 86% of children carried pneumococci in the nasopharynx at least once. The mean prevalence of carriage was 36% in children under two years of age (26).

However, within high income countries some communities have been found to have an increased risk of carriage. Studies of Native American and Native Alaskan communities in the United States have shown increased prevalence of carriage in all age groups (28). A study of White Mountain Apache and Navajo children in the United States found that 92% of 410 children surveyed acquired carriage of pneumococcus at least once during the study period, while the mean prevalence of carriage in children under two was 75% (29). This mean prevalence of carriage is two to three times higher than studies of similarly aged non-Native American children in the United States (30).

iii. Biology and risk factors of disease

Compared to the prevalence of asymptomatic carriage, pneumococcal disease is a rare consequence of acquisition of pneumococcal carriage (31). Pneumococci can cause symptomatic infections of the mucosa of the auditory canal (otitis media), the upper respiratory tract (sinusitis) or the lower respiratory tract (pneumonia) (24). Pneumococcal infection can also result in invasive pneumococcal disease (IPD), defined as isolation of pneumococci from blood, cerebrospinal fluid (CSF), or other normally sterile sites (32). In contrast with asymptomatic carriage, the polysaccharide capsule is known to be a key virulence factor which is associated with disease. The majority of pneumococci isolated from the blood or lungs are opaque phase organisms with a thick capsule and such organisms have been shown to be more efficient at causing invasive disease in rodents (33). The denser capsules of the opaque state provide greater protection against opsonisation and killing through phagocytosis (17).

Risk factors for pneumococcal disease are increased age, underlying chronic illness in adults (34,35). In children day care centre attendance is a risk factor for disease, as day care attendance increases the risk of carriage and thus the risk of disease (36). Serotype is also an important risk factor for
disease. Some serotypes have been shown to be more likely to cause disease than others (31). For example, in the United States seven serotypes, 4, 6B, 9V, 14, 18C, 19F, and 23F, were shown to be responsible for over 70% of the cases of IPD in the pre-vaccine era (37). In a study of IPD and carriage the authors showed significant differences in the rate of IPD per carriage acquisition. Serotypes 4, 14, 7F, 9V and 18C were associated with rates of 120 IPD cases per 100,000 acquisitions of carriage, and serotypes 23F 6A, 16F, 6B and 15B/C were associated with < 10 IPD cases per 100,000 acquisitions (31).

1.2 Interventions
1.2.1 Vaccination

Two types of vaccine against pneumococcal disease are currently licensed for use, a 23 valent polysaccharide vaccine, and multi-valent conjugate capsular polysaccharide vaccines. The 23 valent polysaccharide vaccine was licensed for use in 1983 and provides protection from invasive disease caused by serotypes 1, 2, 3, 4, 5, 6B, 7F, 8, 9N, 9V, 10A, 11A, 12F, 14, 15B, 17F, 18C, 19A, 19F, 20, 22F, 23F, and 33F (37). Vaccine effectiveness in preventing IPD caused by the constituent vaccine serotypes has been shown to be 77% in adults over 60 (38,39). While there is evidence that this vaccine is effective in preventing disease in the elderly and adults with chronic disease, the vaccine does not protect children from IPD. Pneumococcal polysaccharide capsules do not elicit a strong immune response in children (40), however by conjugating the polysaccharide capsule to a carrier protein a stronger immune response can be achieved. Thus, the pneumococcal vaccine currently used to prevent IPD in children is the seven valent conjugate polysaccharide vaccine (PCV7) (40). This vaccine will be replaced by a higher valent conjugate vaccine in the near future (2,41).

i. Impact of the polysaccharide conjugate vaccine

PCV7 provides protection from IPD caused by seven serotypes, 4, 14, 6B, 19F, 18C, 23F and 9V. Antibodies are raised against each of the polysaccharide capsules and so the vaccine is serotype specific (42). These serotypes were chosen for inclusion in the vaccine as together they were responsible for the majority of IPD in the United States prior to the year 2000 (40). In randomized control trials the vaccine was shown to be 98% effective in preventing IPD caused by the seven constituent serotypes (42). In the United States the rate of disease caused by any vaccine serotype was 156 per 100,000 in children under the age of two in 1998-1999, after vaccination was introduced this rate fell to 33.6 per 100,000 children in 2001 (43). In the United States the introduction of PCV7 has resulted in a net reduction of IPD in all age groups, both in children vaccinated and in older adults. In all age groups IPD caused by all serotypes dropped from an average of 24.3 cases per 100,000 persons in 1998 and 1999 to 17.3 per 100,000 in 2001. In children
under the age of two the rate of disease caused by all serotypes declined from 188 cases per 100,000 in 1998 and 1999 to 59 per 100,000 in 2001 (P<0.001) (43).

The decline in IPD in unvaccinated groups is due to indirect or herd effects, as vaccination with PCV7 has led to a decline in nasopharyngeal carriage of the seven vaccine serotypes in vaccinated children (44). A community randomised trial has shown that unvaccinated siblings of vaccinated children are less likely to be colonised with vaccine serotypes (45). As the highest levels of carriage prevalence and transmission are found in young children, PCV7 vaccination has resulted in a decline in disease caused by PCV7 serotypes in all age groups because the majority of transmission events in older age groups is assumed to occur due to contact between young children and older individuals (41).

Vaccination has not resulted in a decline in asymptomatic carriage in children, rather PCV7 vaccination leads to a reduction in nasopharyngeal carriage of vaccine serotypes (VT) while creating immunological and consequently selective pressure for an increased prevalence of non-vaccine serotypes (NVT). A consequence of this has been serotype replacement (41). Serotype replacement in nasopharyngeal carriage has been demonstrated in a series of studies of children in Massachusetts communities. The authors collected samples from nasopharyngeal carriage in the period before and after the introduction of vaccination and demonstrated a significant decline in the prevalence of VT serotypes and concomitant increase in the prevalence of NVT serotypes in carriage (46-48). Invasive disease among children and adults caused by NVT serotypes has significantly increased in the United States post-vaccination, and the majority of IPD among children under five years of age is now due to non-vaccine serotype 19A (49,50).

Recombination plays an important role in allowing the survival of genotypes associated with vaccine serotypes. In the wake of PCV7 vaccination there was concern that acquisition of an NVT capsule through recombination by a pneumococcal strain capable of causing invasive disease would negatively impact on the effectiveness of any serotype-specific vaccine (51). Immunization by PCV7 results in immunity against the capsular serotype, therefore genotypes associated with vaccine serotypes can survive in a vaccinated community though capsular switching. This is when the genes encoding one type of polysaccharide capsule are replaced, via transformation and recombination, with the genes encoding a different type of capsule (51). Capsular serotypes have a stable association with distinct genotype (52), however capsular switching from one serotype to another has previously been demonstrated (53-55). Capsular switching has been demonstrated in serotype 19A (an NVT) after the implementation of PCV7 vaccination. Researchers have shown that a genotype previously associated with serotype 4 (a VT) has been found associated with the NVT capsule 19A (49,51,56). Authors have characterised these strains and demonstrated that the
association of serotype 19A capsule with the serotype 4 genotype arose as a result of a putative recombination event in 2003 (51).

1.2.2 Antibiotic treatment and resistance

The first line treatments for pneumococcal disease are β-lactam antibiotics such as penicillin. When such treatment fails the recommended second line treatments are third generation cephalosporins and vancomycin (57,58).

i. Mechanism of resistance

β-lactam antibiotics inhibit bacterial growth through binding to the cell wall synthesising enzymes, penicillin binding proteins (PBP), and therefore interfere with the synthesis of the pneumococcal cell wall during growth and division. Pneumococci develop resistance to β-lactam antibiotics through alterations to the PBPs which reduce their affinity to this class of antibiotics (59). Six PBPs have been described in the pneumococcus: 1a, 1B, 2x, 2a, 2b and 3 (60). Resistant strains show a mosaic structure in genes encoding 2b, 2x and 1a PBPs (61-63). A mosaic gene structure is where genes are constituted of regions identical to a susceptible gene interwoven with regions with a divergent nucleotide structure (61). Highly divergent regions within the PBP genes of the resistant pneumococci contrast with the uniformity of the nucleotide sequences of other genes, and with the uniformity of the PBP genes in penicillin susceptible pneumococci (60). Thus, it is thought that altered PBP genes have arisen by localized interspecies recombination, involving the PBP genes of closely related streptococci (12,61).

PBP alterations occur in a stepwise fashion, which confer different degrees of resistance to the antibiotic class. Pneumococci with no PBP alterations are fully susceptible to β-lactam antibiotics, primary alterations in the PBPs result in reduced susceptibility, which through further alterations can progress to intermediate susceptibility, and finally full clinical resistance (59,64). Although intermediately susceptible strains are less susceptible to antibiotics, diseases caused by such strains are successfully treated through increasing the dosage of β-lactam antibiotics, with the exception of meningitis (65).

For individuals where treatment with β-lactam antibiotics is not possible (for example due to allergy), antibiotics from the macrolide class such as erythromycin are used. Although chemically distinct, macrolide, lincosamine and streptogramin antibiotics all bind to a ribosomal subunit and act as protein synthesis inhibitors (59). A number of mechanisms which confer resistance to these antibiotics have been described. For example the acquisition of an erythromycin ribosomal methylation gene (erm) results in a posttranscriptional modification of the 23S subunit of ribosomal
RNA, which blocks the binding of the macrolide to the ribosome. Expression of the gene results in high-level resistance to macrolide, lincosamines and streptogramin B (66). Another mechanism of resistance is through acquisition of genes encoding an efflux pump. Acquisition and expression of the mefA gene allows the pneumococcus to effectively excrete the antimicrobial reducing intracellular erythromycin, azithromycin and clarithromycin to ineffective concentrations (67). In contrast to β-lactam resistance, macrolide resistance via these mechanisms is absolute, and cannot be overcome by increasing the dosages of antibiotics (59,68,69). Macrolide resistance determinants are found on mobile genetic elements (transposons), which unlike PBP genes are not a normal component of the bacterial genome (70).

ii. Spread of antibiotic resistance

An antibiotic resistant phenotype can spread in the host population through transmission. For example many multi-drug resistant strains of 23F isolated from different countries have related genotypes, identical mosaic PBP genes and the same serotype suggesting global clonal spread (71). In section 1.1.2 I briefly referred to the fact that recombination was an important factor in the creation and spread of antibiotic resistance in pneumococci. I have already discussed the role of recombination in creating mosaic genes that confer penicillin resistance, but recombination also allows pneumococci to spread resistance horizontally through the population. When homologous recombination transfers a gene from one genotype to another, the PBP gene does not change. Thus it can be assumed that identical PBP genes found in multiple genotype or serotypes have spread horizontally through recombination rather than arising independently multiple times (72).

Research into the evolution of antibiotic resistance has largely focused on the influence of treatment on prevalence of resistance. Cohort studies of invasive disease have shown that antibiotic use increases the risk that the patient will be infected with a resistant strain of pneumococcus (36). At a national level there is a remarkable correlation between antibiotic use in a nation and prevalence of pneumococcal resistance. The ESAC study calculated outpatient antibiotic consumption in Europe between 1997 and 2003. This study showed wide variation in antibiotic use in different countries and seasonal trends in usage. When the total consumption of all antibiotics was compared between countries, consumption was found to vary by a factor of 3 between the country with the highest usage (31.4 daily doses per 1000 inhabitants per day in Greece) and the country with the lowest (9.8 daily doses per 1000 inhabitants per day in the Netherlands) (73). In a study of 26 European countries Goossens et al. (74) showed that the prevalence of resistance in invasive disease and found a significant association between penicillin use and prevalence of penicillin resistance in IPD (74).
iii. Cost of resistance

Antibiotic resistance is often assumed to be associated with reduced fitness (the capability of a genotype to survive and reproduce) (75). This cost may arise because additional DNA is synthesized by the resistant bacteria when resistance is encoded on plasmids or transposons. There is also an energetic cost to the bacterium from the maintenance of antibiotic transport mechanisms such as efflux pumps. It is also possible that changes in the PBPs of resistant pneumococci have a negative effect on their function, thus decreasing the fitness of resistant pneumococci (70). If resistance is associated with reduced fitness, it has been proposed that a reduction in antibiotic use would benefit fitter susceptible strains of pneumococci enabling them to outcompete resistant strains over time (75). Estimation of the cost of antibiotic resistance in pneumococci has to date examined the fitness of resistance strains in vitro and in mouse models of infection. For example a study which investigated the relative fitness of gemifloxin resistant strains found that resistant strains had an increased generation time when compared to susceptible strains (76). To date there has been no examination as to whether resistant pneumococci have reduced fitness in their natural habitat (human nasopharynx). However promising research has shown that a dramatic reduction in antibiotic use results in an associated reduction in resistant pneumococci, suggesting that resistance is associated with a fitness cost and consequently a reduction of community antibiotic use can result in the out competition of resistant strains by sensitive strains (77).

It should be noted that the interaction between antibiotic resistance, fitness and virulence can be more complex. In some bacterial species, there is evidence that ‘compensatory mutations’ can restore wild-type fitness to strains which carry antibiotic resistance (78,79). Much of the research into the effect of antibiotic resistance on fitness has been restricted to the laboratory, there is a real need to examine the impact (or lack of impact) of antibiotic resistance on fitness in the human nasopharynx.

1.2.3 Mathematical models of interventions

i. Mathematical models of antimicrobial resistance

Previous research into the spread of antibiotic resistance has examined the role of different factors that influence the prevalence of antibiotic resistance in bacterial populations. Both generalised and pathogen specific epidemiological mathematical models of drug resistance have been published which examine community and hospital acquired infections. In each model the major selective pressure driving increasing prevalence of antibiotic resistance is the volume of drug use (80).
Studies such as these present the patterns of emergence and establishment of drug resistance in clinical community settings as being similar and consistent. After the introduction of antibiotic treatment a long period of very low prevalence of resistance is followed by a phase of rapid increase and the slow approach to an equilibrium level <100% (80-83). Studies which are most appropriate to understand the evolution of resistance in pneumococcal populations are those that examine the spread of resistance in the community. The emergence and establishment of penicillin resistance in pneumococci in Iceland has been characterised in this way (82). The authors simulated the evolution of antibiotic resistance and constructed a population genetic framework embedded in a transmission model which showed that penicillin resistance increased over time in a community following the form of a sigmoid curve. The authors then fitted the model to the prevalence of resistance in pneumococcal infection in Iceland over a number of years. During the 1980s the first pneumococcal resistant isolates were identified in Iceland and the frequency of resistance rose rapidly. In response, antibiotic consumption by children was reduced by 12.7% between 1992 and 1995. As a result of the intervention, resistance peaked in 1993 at 19.8% and then declined (82). These mathematical models quantified the relationship between resistance and treatment rates; they identified a threshold treatment rate over which the prevalence of resistance increases rapidly to the equilibrium level (81-83).

More recent research has explored why sensitive strains can persist in pneumococcal populations, even under constant treatment over a long period of time. An unusual characteristic of the impact of antibiotic use in the pneumococcal population has been the maintenance of antibiotic sensitive strains over a long period of time, this runs counter to predictions from simulations of the spread of antibiotic resistance in most mathematical models. In the study of penicillin resistance in Iceland, peak prevalence of resistance was 19.8%. Austin et al. (80-83) estimated that equilibrium penicillin resistance in this population would be ~20% if there was no reduction in antibiotic use. The authors overcame the problem of explaining coexistence in a mathematical model by explicitly incorporating strain difference between resistant and sensitive strains. The sensitive strain was less transmissible than the resistant strain (even in the absence of antimicrobial use), but better at superinfection (80-83). However there is no empirical evidence to support this assumption. Thus, although Austin et al. (80-83) successfully fitted the mathematical model to data, the assumptions that underlie the model structure are unproven and the problem of explaining the mechanism of co-existence remains unsolved. This problem is significant as it undermines mathematical model based predictions in public health management and highlights a fundamental gap in our understanding of the dynamics of antimicrobial resistance.
Lipsitch et al. developed a suite of mathematical models of the pneumococcal system to examine the conditions under which sensitive strains can be maintained in the population (84,85). The initial model considered the scenario of single carriage in which the hosts may carry one strain or the other. This model was unsatisfactory as it predicted that the sensitive or resistant strain would be competitively excluded and so does not adequately describe the pneumococcal population (86). A second model allowed dual carriage but predicted a stable equilibrium even in the absence of antibiotic treatment or when the competing strains were indistinguishable in fitness, which is clearly not a realistic result (87).

ii. Mathematical model of vaccine impact

An integral part of simulating the impact of vaccination on the pneumococcal population is simulating the stable coexistence of serotypes before and after vaccination. To date a satisfactory mathematical model which achieves this has not been developed. This is largely due to the fact that the mechanism maintaining many serotypes in the population is not understood. Different serotypes have many biological differences (88), these include, duration of carriage, invasiveness, accessibility of surface determinants (89), and antimicrobial resistance (90). What isn’t clear is which if any of these mechanisms are responsible for the maintenance of multiple serotypes in the population over time (84).

In the absence of this information a model with an explicit coexistence mechanism can only produce tentative predictions as to the impact of vaccination on serotype prevalence and coexistence. However such models can provide a useful tool to understand the impact of vaccination on a multi strain system (84). The majority of mathematical models which consider vaccination in a multi-strain system have considered viral infections (91-94). The dynamics of multiple strain infection, re-infection and transmission may be very different for a bacterial system given the differing immune reactions to the two organisms. Lipsitch considers the impact of a serotype specific pneumococcal vaccine using a mathematical model of a bacterial multi strain system. Using the model, the author showed that in a two serotype system the reduction in carriage from the vaccinated serotype will be exceeded by an increase in carriage from the non-vaccine serotype (86). Models of multi strain systems predict the competitive exclusion of vaccine serotypes in the population—a prediction that has been largely borne out by the changes in serotype prevalence in carriage (48).

1.3 Typing pneumococci

To understand the clinical impact, epidemiology, population and evolutionary biology of pneumococcal populations it is vital to index the variation between strains of pneumococci. Broadly
two distinct methods are currently utilised. The first identifies antigenic variation in the pneumococcal population and the second examines other genetic variation.

1.3.1 Antigenic typing

As I have discussed in previous sections the capsular polysaccharide of the pneumococcus is an important identifier which is associated with the likelihood of disease, antibiotic resistance or vaccine protection. A typing method which identifies different serotypes is therefore clinically important. Antigenic typing mechanisms identify the reaction of antibodies to a bacterial antigen. In the case of the pneumococcus this is the capsular polysaccharide (95).

Currently the most common method of determining the serotype of a pneumococcal isolate is the quellung reaction. This is a biochemical reaction in which anticapsular antibodies bind to the capsule of a bacterium together with India ink, causing the capsule to swell and its outlines to become clearly demarcated. The capsule can then be identified through microscopy (96). For certain common serotypes this method has been advanced so that multiple serotypes can be identified in a semi automated fashion using a multiplex immunoassay with distinct labelled microspheres conjugated to serotype specific antibodies (97). Although the quelling reaction is the gold standard for serotype discrimination it does have disadvantages. The high cost of antisera, the requirements for technical skills and the complexity in interpretation limit its ease of use in all settings (98). PCR techniques to identify the capsular polysaccharide have been developed (99,95), however these have not been developed for comprehensive surveillance and so, to date, the standard technique used in the surveillance of pneumococcal isolates from disease or carriage is the quelling reaction.

Antigenic typing of pneumococci has been used to identify the variation in serotypes causing invasive disease. This has allowed capsular polysaccharide vaccines to be designed so that they maximise their impact against serotypes which most commonly cause invasive disease. In addition surveillance of the serotypes causing invasive disease allows us to examine the impact of vaccination campaigns.

1.3.2 Molecular typing

As I have discussed in previous sections, the identification of capsular switching and the transfer of antibiotic resistant genes between serotypes has highlighted the importance of identifying pneumococcal genetic diversity within a serotype. To do so a tool is needed to provide a suitable level of discrimination to distinguish between pneumococcal isolates of the same serotype.
Multi Locus Sequence Typing (MLST) was developed as a tool for clinical microbiologists and epidemiologists who require a suitable typing mechanism to compare strains of interest from different regions, countries and over time (100). This requires a typing scheme with a high level of discrimination so that all isolates assigned to a type are likely to have shared a recent common ancestor and isolates from a more distant common ancestor are not assigned to the same type (101).

Global and long term epidemiological research requires a typing mechanism that characterises regions of the genome where variation is accumulating very slowly, is likely to be selectively neutral (or not subject to frequent convergent selection leading to homoplasies), and where the mechanism underlying the generation of variation is well understood (102). For this reason MLST identifies sequence variation in housekeeping genes which encode essential metabolic functions. Compared to nucleotide sequencing of a gene under selection only a small number of alleles at each locus can be identified using this method, however high level discrimination between strains is achieved by characterising variation at multiple loci (100).

The MLST scheme used to identify a pneumococcal strain characterises sequence variation at seven housekeeping genes: *ddl*, *spi*, *recP*, *aroE*, *gdh*, *gdh*, *gki* and *xp* (103). Alleles are identified by obtaining the nucleotide sequences of internal fragments of each gene. The gene fragments are amplified from chromosomal DNA using PCR with appropriate primers and then sequenced. For each gene fragment the nucleotide sequences are then compared to all previously identified sequences at that locus and assigned an allele number. Newly identified sequences are assigned a new allele number and isolates with identical sequences are assigned the same allele number. For each strain the combination of seven allele numbers defines its sequence type (ST), this profile is then added to the mlst.net internet database (100).

The first benefit of MLST is that high discrimination can be achieved when characterising a strain, while allowing researchers to identify lineages of isolates over a long period of time or wide geographic area (100). The second is that a multi locus typing scheme is particularly appropriate to examine pneumococcal strains where recombination plays an important role in the generation of genetic variation (104). Homologous recombination can obscure the relationship between two closely related strains and lead to inconsistent relationships between strains inferred through sequencing a single gene. The use of MLST means that relationships identified between strains are more robust; if the alleles at the majority of loci in two strains are identical, a recombination event at one locus does not obscure the relationship between two strains (105). MLST has proven extremely useful in examining the population genetics of pneumococcal populations (106,105).
has allowed researchers to identify globally distributed antibiotic resistant clones of pneumococcus (107,102) but also gain a greater understanding of the evolution of strains in asymptomatic carriage (108).

1.4 Aims of thesis

The overall aim of this thesis is to examine the impact and evolution of antibiotic resistance and the impact of strong selective pressure on the population structure of *S. pneumoniae*.

To achieve this overall aim I will examine four questions in the course of this thesis:

- **What is the added burden of mortality associated with antibiotic resistant pneumococci?**

  Research into the burden of pneumococcal disease has shown that pneumococcal pneumonia, meningitis and bacteraemia are responsible for a significant burden of mortality in children under five. Surveillance of the prevalence of penicillin resistance has shown that in some countries more than 50% of isolates are resistant to penicillin. In Chapter 2 I will carry out a systematic review and meta-analysis to examine whether antibiotic resistance (primarily penicillin resistance) contributes an added burden of mortality.

- **How and why are antibiotic resistant strains distributed among serotypes?**

  Even if resistance to antibiotics in the pneumococcus does not have an obvious impact on mortality in invasive disease at this time, treatment failure in non-invasive and invasive disease leads to the use of antibiotics which are more expensive to administer. It is important then to study and understand the distribution of antibiotic resistance in pneumococcal populations, and what gives rise to it. Given that the serotypes of the pneumococcus have distinct characteristics and antibiotic resistance is known to associate more frequently with some serotypes than others I will use a mathematical model and analysis of serotype prevalence to examine how and why antibiotic resistance is distributed among the serotypes. This will be described in Chapters 3 and 4.

- **What effect does the spread of advantageous alleles have on genetic diversity in the population?**

  The growing collection of MLST data provides an opportunity to understand how the spread of evolutionary advantageous alleles affects the genetic diversity in the pneumococcal population. I will carry out a review of the bacterial and eukaryotic population genetics literature to examine whether current techniques have adequately described the impact of selection on a bacterial population like
S. pneumoniae where homologous recombination plays an important role in generating and maintaining diversity. Chapter 5 will detail the results of the review and provide a comparison between a model incorporating eukaryotic type recombination and one incorporating bacterial type recombination. I shall construct and simulate a stochastic model of a selective sweep of advantageous alleles through a bacterial population. This research will be reported in Chapter 6.

- **How does vaccination impact on diversity assessed by MLST in the pneumococcal population?**

Use of the pneumococcal conjugate vaccine has significantly changed the distribution of serotypes in disease and carriage. In Chapter 7 I will take insights from the mathematical model of a simple selective sweep and analyse whether the predictions of this model can adequately describe the selection in the pneumococcal population which arose as a result of vaccination.
2 The impact of pneumococcal antibiotic resistance on mortality: a review and meta-analysis

2.1 Introduction

The threat from antimicrobial resistance has long been illustrated as an arms race between the development of effective antibiotic therapy and the evolution of drug resistance. Antimicrobial resistant pathogens are generally thought to result in infections with higher morbidity and mortality than similar infections with antibiotic sensitive pathogens. This increased negative clinical impact is most likely due to a lack of suitable treatment, as a delay in the treatment of the infection can be detrimental to the patient. As penicillin resistant pneumococcal infection can still be successfully treated with a number of antibiotics, an infection with a penicillin resistant organism is not necessarily characterised as clinically costly at the current time (109). However, given that resistant pneumococci are common (110), it is important to estimate whether resistance already adversely impacts on a patient’s survival.

Pneumococcal infections are classed as susceptible to an antibiotic when the recommended therapeutic dosage will successfully clear an infection, intermediately resistant when a higher dosage of the same antibiotic can still successfully clear an infection and resistant to an antibiotic when no dosage of the antibiotic can clear an infection. Several studies have investigated the differences in mortality rates of infections with antibiotic susceptible and non-susceptible (intermediately and fully resistant organisms) pneumococci, to assess the impact of antibiotic resistance in pneumococcal infections. These studies have not produced consistent results. Some cohort and case-control studies have suggested that there is no significant impact of pneumococcal non-susceptibility on mortality (18). However a meta-analysis of the impact of penicillin non-susceptible pneumococcal (PNSP) pneumonia found that there was a significant increased risk of mortality associated with PNSP infections (109). Some authors have criticised this interpretation (110,111), pointing out that pneumococcal pneumonia was treated with antibiotics other than penicillin in the majority of cases studied. Furthermore, when a patient was recorded as receiving discordant therapy (when a patient is treated with penicillin although the infection is caused by PNSP pneumococci) there was no increased adverse effect when compared to patients treated with a concordant therapy (109,111). They suggest that any association between PNSP pneumonia and increased mortality is a result of confounders associated with PNSP infections and not due to resistance itself (112,113).
Such confounders make it difficult to distinguish between: a) an association between antibiotic resistance and poor outcome and b) causality between resistance and poor outcome. Possible confounders include age, underlying disease and severity of disease. A patient who has an underlying medical condition or is of advanced age may have a greater risk of death from a pneumococcal infection than a younger individual with no underlying medical condition (114). A patient with an underlying condition may also be more likely to previously be prescribed antibiotics (for any infection, not necessarily pneumococcal). As prior antibiotic therapy is associated with an increased risk of colonisation with a resistant pneumococcal strain (OR 4.0) (115), such a patient is more likely to carry asymptomatically, or suffer from a symptomatic infection, with a resistant strain.

Assessing the impact of PNSP and resistant infection in pneumococcal disease is further complicated by the range of pneumococcal disease and the ability of antibiotics to penetrate the site of infections differently. For example, it is known that penicillin penetrates the cerebral spinal fluid (CSF) poorly, so much higher dosages of penicillin are needed to be effective in the treatment of meningitis compared to pneumonia or bacteraemia (116). Thus, it is important then to examine each invasive disease syndrome separately.

As I mentioned in the above, a meta-analysis which examined pneumococcal pneumonia was published, however no comparable study on meningitis or bacteraemia has been published. In recent years the possible impact of penicillin non-susceptibility on meningitis has been highlighted by Klugman et al. (112), who noted that penicillin is often the treatment of choice for meningitis in Sub-Saharan Africa (112,116). There is a need therefore to carry out an analysis of the impact of penicillin non-susceptibility in pneumococcal meningitis.

2.1.1 Aims

In order to estimate the impact of resistance on IPD I will perform a systematic review and meta-analysis. I will examine the impact of resistance on mortality as mortality is an outcome that can be easily compared across studies and is most widely reported. Given that treatment of pneumococcal infection has changed in developed countries in response to increasing prevalence of resistance, I hypothesise that the impact of resistance on mortality will be most keenly felt in those countries whose routine treatments have not changed in response to resistance.

To examine this I will:

- Carry out a systematic review of all IPD,
• Examine the impact of resistance on mortality in three major syndromes: meningitis, bacteraemia and invasive pneumonia,
• Attempt to examine the impact of known confounders.
2.2 Method

2.2.1 Literature search terms and strategy

The search terms used in the systematic search of PubMed are outlined in Table 2.1. The search strategy was made up of four sections: A) Organism B) Study type, C) Resistance, D) Outcome and E) Exclusions. The search was conducted on papers published up to March 2009. When a group of studies was collected from the systematic search results, further studies were identified through reading the reference lists of collected papers.

Where appropriate I used PubMed MeSH terms which increase the breadth and success of the search. MeSH is the United States National Library of Medicine’s controlled vocabulary used for indexing articles for PubMed/MEDLINE. MeSH terms are assigned by hand to journal articles submitted to PubMed. They provide a consistent way to retrieve information where different terminology may be used to describe the same topics or concepts (117). For example using *Streptococcus pneumoniae* as a MeSH term will return articles where “*Streptococcus pneumoniae*”, “*S. pneumoniae*”, “pneumococcus” etc. are MeSH subheadings that have been assigned to the text. In contrast searching for “*Streptococcus pneumoniae*” as a conventional search term will return only articles where the phrase *Streptococcus pneumoniae* is present in the text.
Table 2.1 Search Terms

<table>
<thead>
<tr>
<th>Search component</th>
<th>Search Terms</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AND Drug resistance</strong></td>
<td>&quot;Drug Resistance&quot;[MeSH] OR &quot;Drug Resistance, Multiple, Bacterial&quot;[MeSH] OR &quot;Drug Resistance, Bacterial&quot;[MeSH]</td>
<td>1102</td>
</tr>
<tr>
<td><strong>AND Outcome</strong></td>
<td>mortality OR &quot;short term mortality&quot;</td>
<td>354</td>
</tr>
<tr>
<td><strong>AND Exclusion</strong></td>
<td>&quot;Pneumococcal Vaccines&quot;[MeSH Major Topic]</td>
<td>349</td>
</tr>
</tbody>
</table>

**i) Inclusion criteria**

Studies were eligible for inclusion if they reported or allowed calculation of the mortality for antibiotic resistant, non susceptible and sensitive pneumococcal infections. All studies of IPD were included as long as the study identified that the infection was caused by Streptococcus pneumoniae and death was a result of the pneumococcal infection. I also included cohort and case control studies, both prospective and retrospective, and infections from all age and patient groups.

**ii) Screening and article retrieval**

To determine whether articles were eligible for further review the title and abstract of the article was read at the search stage. If not relevant the article was screened out at this stage. Articles that only contained the following types of data listed below were excluded at this stage: case reports, vaccine safety studies, nasopharyngeal carriage studies, animal model studies, studies of species
other than *S. pneumoniae*. If an article was eligible for further review it was retrieved electronically and when this was not possible a hard copy was obtained from Imperial College Library.

**iii) Exclusion criteria and variables recorded**

I reviewed the whole text of the retrieved articles and excluded those that did not fit the pre-defined inclusion criteria. Articles were excluded if they did not report *S. pneumoniae* specific rates of disease or mortality. Studies which lacked adequate data to form a contingency table describing survival and antibiotic specific resistant disease status were also excluded.

**iv) Location and year of study**

The country or countries included in the study was recorded. If a study was carried out in more than one country but the authors did not report separate results for each country, all country names were noted and the reported data were recorded. I noted the year of publication of the article and the time period during which the study was undertaken.

**v) Study design and population**

I recorded whether data collection was prospective or retrospective, whether the study could be classified as case control or a cohort study and what statistical model had been used to analyse data. I also noted the age range and size of the study population and the disease inclusion criteria and antibiotic resistance tested in the study.

**vi) Study outcomes**

The principal outcome of interest recorded was mortality. I noted whether the mortality definition recorded in the study was “in hospital mortality”, “30 day mortality” or all mortality. In order to calculate contingency tables for mortality I recorded the number of cases and number of deaths for study participants with a resistant or susceptible infection. Where available I also recorded the length of stay in hospital and costs associated with treatment for resistant and susceptible infections. I also collected data on the antibiotic treatment used where this was stated. Where an adjusted mortality Odds Ratio (OR) had been calculated in a study using a statistical model this was recorded.

**vii) Antibiotic resistance**

I recorded what antibiotic resistance had been tested for and the susceptibility breakpoints that had been used in testing.
2.2.2 Statistical methods

The initial meta-analysis was based on the unadjusted OR calculated from raw deaths and survival data collected from the studies. ORs and 95% confidence intervals for individual studies were calculated and pooled using the Review Manager software package (RevMan 5.0) (118). A fixed effects model was used to pool the effect estimates. Crude ORs were weighted using the Mantel-Haenszel weighting method. Forest and funnel plots based on the data were formed using the RevMan software. All sub-group analysis was also carried out using RevMan software.

When calculating ORs for penicillin susceptibility I compared the number of deaths among patients with a PNSP infection to deaths among patients with penicillin sensitive *S. pneumoniae* (PSP) infections. In line with the pre-2008 (Clinical and Laboratory Standards Institute) CLSI MICs, PSP isolates were isolates with a minimum inhibitory concentration (MIC) of ≤0.06 µg/mL and PNSP isolates were those with a MIC of over 0.1 µg/mL. For isolates tested for cefotaxime resistance, I compared the number of deaths in the group infected with isolates with an MIC over 0.25µg/ml to those with an MIC below this. I calculated separate ORs for each resistance type. In these analyses I only considered “in hospital” deaths or deaths that occurred thirty days after diagnosis with *S. pneumoniae* infection. I calculated pooled estimates of adjusted OR by combining the separate estimates of inverse variance-weighted log OR from each study. I carried out subgroup analysis where appropriate. I calculated pooled ORs for each syndrome and study country.

2.3 Results

2.3.1 Search results

The systematic literature search was carried out until 30th March 2009. The search generated 1102 results when restricted to those that reported mortality data and studies which carried the MeSH term “Pneumococcal Vaccine” as a major topic the number of results generated was 349. I reviewed the title and abstract of these studies and collected 65 studies suitable for further review. From the review of the full text of these papers I excluded 35 papers. References of excluded studies and their reasons for exclusion are presented in Table 2.2. The most common reason for exclusion at this stage was that the study did not have resistance related mortality data. Other studies were excluded as they were serotype surveillance studies and reviews with no original data.

Thirty papers were selected that had data suitable for inclusion in the review and I also included ten suitable papers from review of the bibliographies of systematically searched papers. A summary of the characteristics of these studies is presented in Table 2.3. The most common study country was the United States where thirteen studies were carried out. The earliest study started in 1979 and the
most recent study finished in 2005. Included in the meta-analysis were ten bacteraemia studies, eight studies of meningitis, twelve studies of pneumococcal pneumoniae, one study of pneumococcal peritonitis, one study of endocarditis and nine studies of mixed invasive pneumococcal infection (IPnI). The largest study size had 3452 patients with PSP infection and 741 patients infected with PNSP infection. The smallest study had 64 patients in total.

The most common antibiotic susceptibility analysed was penicillin non-susceptibility. The other antibiotic studied was cefotaxime. When the initial treatment regime was reported in the study penicillin was the least common form of treatment. Penicillin was reportedly used to treat patients with pneumonia and other IPDs (188) and one study in Hong Kong reported its use in the treatment of bacteraemia (166). Both African studies included, reported that treatment of patients with meningitis was with penicillin (176,112).
<table>
<thead>
<tr>
<th>Author, Year of Publication</th>
<th>Article References</th>
<th>Reason for rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Chomarat et al. 2002), (Echave et al. 2003), (Fernandez Guerrero et al. 2003), (Garcia-Lechuz et al. 2007), (Gentile et al. 2003), (Jette and Lamothe 1989), (Laupland et al. 2004), (Leelarasamee, Dhiraputra, and Hunnangkul 1999), (Levin et al. 2003), (Mattei, Falleiros-Carvalho, and Cavalcante 2008), (Rendi-Wagner et al. 2004), (Siegel 1999), (Watanabe et al. 2000), (Woolfson et al. 1997), (Yanagihara et al. 2004), (Yaro et al. 2006)</td>
<td>(119), (120), (121), (122), (123), (124), (125), (126), (127), (128), (129), (130), (131), (132), (133), (134)</td>
<td>No information on resistance related mortality</td>
</tr>
<tr>
<td>(Chen et al. 2006), (Friedland and Klugman 1992), (Ramos et al. 1996), (Sirinavin et al. 2003), (Sanz Herrero et al. 2003), (Wu et al. 2000)</td>
<td>(135), (136), (137), (138), (139), (140)</td>
<td>Text unavailable</td>
</tr>
<tr>
<td>(Tan 2003), (Tleyjeh et al. 2006)</td>
<td>(141), (142)</td>
<td>Review</td>
</tr>
<tr>
<td>(Ispahani et al. 2004), (Sjostrom et al. 2006)</td>
<td>(143), (32)</td>
<td>Serotype Study</td>
</tr>
<tr>
<td>(Rowland and Turnidge 2000)</td>
<td>(144)</td>
<td>No deaths in the study group</td>
</tr>
<tr>
<td>(Hoffman et al. 2003)</td>
<td>(67)</td>
<td>Mouse Model</td>
</tr>
<tr>
<td>(Wandi, Kiagi, and Duke 2005)</td>
<td>(145)</td>
<td>Mixed Bacterial Study</td>
</tr>
<tr>
<td>(Huang et al. 2007)</td>
<td>(146)</td>
<td>Incorrect Species</td>
</tr>
<tr>
<td>(Madhi et al. 2000)</td>
<td>(147)</td>
<td>HIV Comparison Only</td>
</tr>
<tr>
<td>(Yu and Baddour 2004)</td>
<td>(148)</td>
<td>Comment</td>
</tr>
<tr>
<td>(Dylewski and Davidson 2006)</td>
<td>(149)</td>
<td>Case Report</td>
</tr>
<tr>
<td>(Sun, Nicolau, and Kuti 2006)</td>
<td>(150)</td>
<td>Economic Analysis</td>
</tr>
<tr>
<td>(Isozumi et al. 2007)</td>
<td>(66)</td>
<td>Genotyping study</td>
</tr>
</tbody>
</table>
Table 2.3 Studies included in meta-analysis

<table>
<thead>
<tr>
<th>Author, Year of Publication</th>
<th>Reference</th>
<th>Country</th>
<th>Study Type</th>
<th>Age(i)</th>
<th>Syndrome</th>
<th>Period of Study</th>
<th>Antibiotic</th>
<th>Initial Treatment Regime</th>
<th>Study Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Ailani et al. 2002)</td>
<td>(151)</td>
<td>USA</td>
<td>Retrospective Case Control</td>
<td>53&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Pneumonia</td>
<td>1993-1999</td>
<td>Cephalosporin</td>
<td>Cephalosporin</td>
<td>78</td>
</tr>
<tr>
<td>(Arditi et al. 1998)</td>
<td>(152)</td>
<td>USA</td>
<td>Prospective Cohort</td>
<td>9&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Meningitis</td>
<td>1993-1996</td>
<td>Penicillin</td>
<td>Cephalosporin</td>
<td>180</td>
</tr>
<tr>
<td>(Aspa et al. 2004a)</td>
<td>(153)</td>
<td>Spain</td>
<td>Prospective Cohort</td>
<td>16 to 97</td>
<td>Pneumonia</td>
<td>1999-2000</td>
<td>Penicillin</td>
<td>...</td>
<td>638</td>
</tr>
<tr>
<td>(Capdevila et al. 2001)</td>
<td>(154)</td>
<td>Spain</td>
<td>Retrospective Case Control</td>
<td>14 to 65</td>
<td>Peritonitis</td>
<td>1979-1998</td>
<td>Penicillin</td>
<td>...</td>
<td>90</td>
</tr>
<tr>
<td>(Castillo et al. 2000)</td>
<td>(155)</td>
<td>USA</td>
<td>Retrospective Cohort</td>
<td>1 to 83</td>
<td>Bacteraemia</td>
<td>1991-1998</td>
<td>Penicillin</td>
<td>...</td>
<td>281</td>
</tr>
<tr>
<td>(Choi and Lee 1998)</td>
<td>(156)</td>
<td>Korea</td>
<td>Retrospective Cohort</td>
<td>0 to 15</td>
<td>IPnI</td>
<td>1985-1996</td>
<td>Penicillin</td>
<td>Penicillin, Vancomycin, Macrolide</td>
<td>72</td>
</tr>
<tr>
<td>(Ewig et al. 1999)</td>
<td>(159)</td>
<td>Spain</td>
<td>Prospective Cohort</td>
<td>26 to 82</td>
<td>Pneumonia</td>
<td>1996-1998</td>
<td>Pen and Ceph</td>
<td>Macrolide</td>
<td>101</td>
</tr>
<tr>
<td>(Falco 2004)</td>
<td>(160)</td>
<td>Spain</td>
<td>Prospective Cohort</td>
<td>62.8&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Pneumonia</td>
<td>1997-2001</td>
<td>Penicillin</td>
<td>Penicillin, Ceftriaxone</td>
<td>247</td>
</tr>
<tr>
<td>(Feikin et al. 2000)</td>
<td>(157)</td>
<td>USA</td>
<td>Prospective Cohort</td>
<td>0 to 85</td>
<td>IPnI&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1995-1997</td>
<td>Penicillin</td>
<td>...</td>
<td>12194</td>
</tr>
<tr>
<td>(Fiore et al. 2000)</td>
<td>(161)</td>
<td>USA</td>
<td>Retrospective Cohort</td>
<td>2 to 94</td>
<td>Meningitis</td>
<td>1994-1996</td>
<td>Cefotaxime</td>
<td>Vancomycin</td>
<td>109</td>
</tr>
<tr>
<td>Author, Year of Publication</td>
<td>Reference</td>
<td>Country</td>
<td>Study Type</td>
<td>Age[^1]</td>
<td>Syndrome</td>
<td>Period of Study</td>
<td>Antibiotic</td>
<td>Initial Treatment Regime</td>
<td>Study Size</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------</td>
<td>---------</td>
<td>------------</td>
<td>---------</td>
<td>----------</td>
<td>----------------</td>
<td>------------</td>
<td>--------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>(Gomez et al. 1995)</td>
<td>(163)</td>
<td>Spain</td>
<td>Prospective Cohort</td>
<td>12 to 84</td>
<td>Bacteraemia</td>
<td>1989-1993</td>
<td>Penicillin</td>
<td>...</td>
<td>71</td>
</tr>
<tr>
<td>(Gomez-Barreto et al. 1999)</td>
<td>(164)</td>
<td>Mexico</td>
<td>Prospective Cohort</td>
<td>0 to 8</td>
<td>Meningitis</td>
<td>1994-1998</td>
<td>Penicillin</td>
<td>...</td>
<td>38</td>
</tr>
<tr>
<td>(Gomez-Barreto et al. 2000)</td>
<td>(165)</td>
<td>Mexico</td>
<td>Prospective Cohort</td>
<td>2 to 4</td>
<td>IPnI</td>
<td>1994-1998</td>
<td>Penicillin</td>
<td>Ceftriaxone</td>
<td>49</td>
</tr>
<tr>
<td>(Ho et al. 2006)</td>
<td>(166)</td>
<td>Hong Kong</td>
<td>Retrospective Cohort</td>
<td>10 to over 70</td>
<td>Bacteraemia</td>
<td>1995-2001</td>
<td>Penicillin</td>
<td>Penicillin</td>
<td>173</td>
</tr>
<tr>
<td>(Iannini et al. 2007)</td>
<td>(167)</td>
<td>USA</td>
<td>Retrospective Cohort</td>
<td>33 to 80</td>
<td>Pneumonia</td>
<td>2004-2005</td>
<td>Erythromycin</td>
<td>...</td>
<td>122</td>
</tr>
<tr>
<td>(J Kellner 1999)</td>
<td>(170)</td>
<td>Canada</td>
<td>Case Control</td>
<td>0 to 15</td>
<td>Meningitis</td>
<td>1991-1999</td>
<td>Penicillin</td>
<td>Ceftriaxone</td>
<td>75</td>
</tr>
<tr>
<td>(Kaplan et al. 1998)</td>
<td>(169)</td>
<td>USA</td>
<td>Prospective Cohort</td>
<td>0 to 18</td>
<td>IPnI</td>
<td>1993-1996</td>
<td>Penicillin</td>
<td>Ceftriaxone</td>
<td>1291</td>
</tr>
<tr>
<td>(Kim et al. 2002)</td>
<td>(34)</td>
<td>South Korea</td>
<td>Retrospective Cohort</td>
<td>18 to 90</td>
<td>Bacteraemia</td>
<td>1996-2001</td>
<td>Penicillin</td>
<td>...</td>
<td>151</td>
</tr>
<tr>
<td>(Kumashi et al. 2005)</td>
<td>(171)</td>
<td>USA</td>
<td>Retrospective Cohort</td>
<td>34 to 72</td>
<td>Bacteraemia</td>
<td>1998-2002</td>
<td>Penicillin</td>
<td>Fluoroquinolones</td>
<td>122</td>
</tr>
<tr>
<td>(Ma, 2002)</td>
<td>(172)</td>
<td>Taiwan</td>
<td>Retrospective Cohort</td>
<td>0 to 14</td>
<td>IPnI</td>
<td>1990-2000</td>
<td>Penicillin</td>
<td>...</td>
<td>72</td>
</tr>
<tr>
<td>(Maddox and Winter 2003)</td>
<td>(173)</td>
<td>UK</td>
<td>Retrospective Cohort</td>
<td>0 to 98</td>
<td>Bacteraemia</td>
<td>1998-2000</td>
<td>Penicillin &amp; Macrolide</td>
<td>Erythromycin</td>
<td>106</td>
</tr>
<tr>
<td>(Martinez et al. 2002)</td>
<td>(158)</td>
<td>Spain</td>
<td>Prospective Cohort</td>
<td>18 to 79</td>
<td>Endocarditis</td>
<td>1978-1998</td>
<td>Penicillin</td>
<td>...</td>
<td>63</td>
</tr>
<tr>
<td>(Maugein et al. 2003)</td>
<td>(174)</td>
<td>France</td>
<td>Survey</td>
<td>0 to over 70</td>
<td>Bacteraemia</td>
<td>1997</td>
<td>Penicillin</td>
<td>...</td>
<td>438</td>
</tr>
<tr>
<td>(Metlay, 2002)</td>
<td>(175)</td>
<td>USA</td>
<td>Retrospective Cohort</td>
<td>53[^3]</td>
<td>Pneumonia</td>
<td>1994</td>
<td>Penicillin</td>
<td>...</td>
<td>192</td>
</tr>
<tr>
<td>(Molyneux et al. 2002)</td>
<td>(176)</td>
<td>Malawi</td>
<td>RCT(Dexamethasone treatment)</td>
<td>2 months-15 years</td>
<td>Meningitis</td>
<td>1997-2001</td>
<td>Penicillin</td>
<td>Penicillin &amp; Chloramphenicol</td>
<td>188</td>
</tr>
<tr>
<td>Author, Year of Publication</td>
<td>Reference</td>
<td>Country</td>
<td>Study Type</td>
<td>Age<a href="i">^</a></td>
<td>Syndrome</td>
<td>Period of Study</td>
<td>Antibiotic</td>
<td>Initial Treatment Regime</td>
<td>Study Size</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------</td>
<td>---------------</td>
<td>---------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------------</td>
<td>------------</td>
<td>--------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>(Mufson, Chan, and Stanek 2007)</td>
<td>(178)</td>
<td>USA</td>
<td>Retrospective Cohort</td>
<td>18 to 96</td>
<td>Pneumonia</td>
<td>1983-2003</td>
<td>Penicillin</td>
<td>Cephalosporin</td>
<td>185</td>
</tr>
<tr>
<td>(Pagliano et al. 2007)</td>
<td>(179)</td>
<td>Italy</td>
<td>Prospective Cohort</td>
<td>0 to 7</td>
<td>Meningitis</td>
<td>1997-2005</td>
<td>Penicillin</td>
<td>Cephalosporin</td>
<td>64</td>
</tr>
<tr>
<td>(Pallares et al. 1995b)</td>
<td>(180)</td>
<td>Spain</td>
<td>Prospective Cohort</td>
<td>19 to 70</td>
<td>IPnI</td>
<td>1994-2000</td>
<td>Penicillin</td>
<td>Ceftriaxone and Penicillin</td>
<td>456</td>
</tr>
<tr>
<td>(Pallares et al. 2002b)</td>
<td>(181)</td>
<td>Spain</td>
<td>Prospective Cohort</td>
<td>19 to over 70</td>
<td>IPnI</td>
<td>1994-2000</td>
<td>Cephalosporin</td>
<td>Ceftriaxone</td>
<td>429</td>
</tr>
<tr>
<td>(Quach et al. 2002)</td>
<td>(182)</td>
<td>Canada</td>
<td>Case control</td>
<td>0 to 18</td>
<td>IPnI</td>
<td>1989-1998</td>
<td>Penicillin</td>
<td>Ceftriaxone</td>
<td>144</td>
</tr>
<tr>
<td>(Sangthawan 2003)</td>
<td>(183)</td>
<td>Thailand</td>
<td>Prospective Cohort</td>
<td>52.5[^b]</td>
<td>Pneumonia</td>
<td>1998-2001</td>
<td>Penicillin</td>
<td>...</td>
<td>46</td>
</tr>
<tr>
<td>(Song 2004)</td>
<td>(184)</td>
<td>9 Asian countries</td>
<td>Prospective Cohort</td>
<td>60.8[^b]</td>
<td>Pneumonia</td>
<td>2000-2001</td>
<td>Penicillin</td>
<td>Penicillin, Cephalosporin</td>
<td>233</td>
</tr>
<tr>
<td>(Turett et al. 1999)</td>
<td>(185)</td>
<td>USA</td>
<td>Retrospective Cohort</td>
<td>0 to 102</td>
<td>Bacteraemia</td>
<td>1992-1996</td>
<td>Penicillin</td>
<td>...</td>
<td>456</td>
</tr>
<tr>
<td>(Wasier et al. 2005)</td>
<td>(186)</td>
<td>France</td>
<td>Retrospective Cohort</td>
<td>0 to 9</td>
<td>Meningitis</td>
<td>1990-2003</td>
<td>Penicillin</td>
<td>Vancomycin</td>
<td>49</td>
</tr>
<tr>
<td>(Winston et al. 1999)</td>
<td>(187)</td>
<td>USA</td>
<td>Case control</td>
<td>30 to 58</td>
<td>IPnI</td>
<td>1994-1996</td>
<td>Penicillin</td>
<td>...</td>
<td>476</td>
</tr>
<tr>
<td>(Yigla 1995)</td>
<td>(188)</td>
<td>Israel</td>
<td>Prospective Cohort</td>
<td>60.2[^b]</td>
<td>Pneumonia</td>
<td>1989-1990</td>
<td>Penicillin</td>
<td>Penicillin</td>
<td>22</td>
</tr>
</tbody>
</table>

i) Study population age is in the form of a range or a) median age and b) mean age

ii) IPnI

iii) Italicized text are studies included through bibliographic rather than systematic searches
2.3.2 Unadjusted ORs

I extracted the number of cases and deaths for antibiotic non-susceptible and susceptible infection. These were inputted into the RevMan software package which was used to calculate individual and pooled ORs. Results were obtained from 3371 non-susceptible infections and 9559 susceptible infections. 580 deaths occurred in the non-susceptible infection group and 1206 deaths occurred in the susceptible infection group. The pooled OR for all studies was 1.30 (CI 1.16-1.46).

Three studies examined the effect of cefotaxime resistance only. Results were obtained from 92 resistant infections and 287 susceptible infections. There were 13 deaths in the group of patients infected with cefotaxime resistant pneumococcus and 48 deaths in the group of patients infected with cefotaxime susceptible pneumococcus. Analysed separately from the PNSP isolates the pooled OR for these three studies is 0.89 (CI 0.45, 1.76).

The remaining studies analysed the effect of PNSP infection on mortality. There were 2753 patients with PNSP infection. 526 of these patients died in hospital or within 30 days of diagnoses. 8500 patients were infected with penicillin susceptible bacteria, 1066 of these patients died. The pooled OR for these studies is 1.49 (CI 1.32, 1.69).

2.3.3 Sub-group analysis

As the effect of cefotaxime resistance was only reported in three studies I restricted further subgroup analysis to penicillin susceptibility only. I grouped the studies by country and syndrome. Shown in Figure 2.1, IPnI and pneumococcal pneumonia had the closest estimated ORs to the average with ORs of 1.27 (CI 1.04 -1.55) and 1.53 (CI 1.28 – 1.84) respectively. Bacteraemia infections had a pooled OR of 1.03 (CI 0.78-1.36). Meningitis infections had a pooled OR of 1.54(CI 1.01-2.34).

I also analysed the studies by country or region where appropriate. This is illustrated in Figure 2.2 which shows the unadjusted ORs for studies carried out in regions or countries. Spain and the United States were analysed distinct from other studies from Europe or the Americas. Asian, other European and North and South American were analysed regionally. The lowest value of pooled OR is 0.67 (CI 0.13-3.41) and was reported by studies carried out in Canada. The pooled OR for the two African studies was 2.25 (CI 1.29-3.94), which was the highest found from the sub regional analysis.

However, any apparent regional effects were insignificant once syndrome and treatment biases were accounted for. Different syndromes were unevenly distributed between regions, for example, there were no studies of pneumococcal meningitis reported from Spain, but there were four
pneumonia studies. There was a small difference between OR for pneumonia from Spanish and United States studies, the pooled OR for pneumococcal pneumonia from United States studies was 1.52 (CI 0.87, 2.66) whereas that for Spanish studies was 1.85 (CI 1.38-2.49), though the difference is not statistically significant. There were only two studies from Africa, one from Malawi and another from South Africa. The study from Malawi is data extracted retrospectively from a randomised control trial of dexamethasone use in the treatment of children with bacterial meningitis (112,176). The study from South Africa was a small prospective cohort of children with pneumococcal meningitis (162). Both these studies reported that patients infected with PNSP meningitis had an increased risk of death compared to a PSP infection.
<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Non-susceptible</th>
<th>Susceptible</th>
<th>Odds Ratio M-H, Fixed, 95% CI</th>
<th>Odds Ratio M-H, Fixed, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1.1 Meningitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arditi 1998</td>
<td>2</td>
<td>35</td>
<td>12</td>
<td>146</td>
</tr>
<tr>
<td>Fiore 2000</td>
<td>2</td>
<td>22</td>
<td>13</td>
<td>87</td>
</tr>
<tr>
<td>Friedland 1992</td>
<td>13</td>
<td>30</td>
<td>11</td>
<td>49</td>
</tr>
<tr>
<td>Gomez-Barreto 1999</td>
<td>1</td>
<td>11</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>Kellner 2002</td>
<td>1</td>
<td>29</td>
<td>3</td>
<td>42</td>
</tr>
<tr>
<td>Molyneux 2002</td>
<td>22</td>
<td>45</td>
<td>45</td>
<td>143</td>
</tr>
<tr>
<td>Paglino 2007</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>Wasier 2005</td>
<td>9</td>
<td>16</td>
<td>14</td>
<td>32</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>188</td>
<td>580</td>
<td>7.0%</td>
<td>1.54 [1.01, 2.34]</td>
</tr>
</tbody>
</table>

Total events 51 103
Heterogeneity: Chi² = 7.13, df = 7 (P = 0.42); I² = 2%
Test for overall effect: Z = 2.01 (P = 0.04)

8.1.2 IPnI

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Events</th>
<th>Total</th>
<th>Weight</th>
<th>Odds Ratio M-H, Fixed, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choi 1998</td>
<td>3</td>
<td>32</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>Feikin 2000</td>
<td>101</td>
<td>383</td>
<td>24.1%</td>
<td>1.26 [1.00, 1.60]</td>
</tr>
<tr>
<td>Gomez-Barreto 2000</td>
<td>11</td>
<td>25</td>
<td>0.8%</td>
<td>1.91 [0.58, 6.23]</td>
</tr>
<tr>
<td>Kaplan 1998</td>
<td>4</td>
<td>221</td>
<td>1062</td>
<td>1.0%</td>
</tr>
<tr>
<td>Ma 2002</td>
<td>10</td>
<td>31</td>
<td>26</td>
<td>0.8%</td>
</tr>
<tr>
<td>Pallares 2002</td>
<td>8</td>
<td>55</td>
<td>130</td>
<td>2.4%</td>
</tr>
<tr>
<td>Quach 2002</td>
<td>1</td>
<td>36</td>
<td>108</td>
<td>0.3%</td>
</tr>
<tr>
<td>Winston 1999</td>
<td>6</td>
<td>65</td>
<td>411</td>
<td>1.6%</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>1206</td>
<td>5253</td>
<td>31.2%</td>
<td>1.27 [1.04, 1.57]</td>
</tr>
</tbody>
</table>

Total events 144 468
Heterogeneity: Chi² = 3.10, df = 7 (P = 0.88); I² = 0%
Test for overall effect: Z = 2.30 (P = 0.02)
### 8.1.3 Pneumonia

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Events</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ailani 2002</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Aspa 2004</td>
<td>42</td>
<td>0.01</td>
</tr>
<tr>
<td>Ewig 1999</td>
<td>8</td>
<td>0.01</td>
</tr>
<tr>
<td>Falco 2004</td>
<td>16</td>
<td>0.01</td>
</tr>
<tr>
<td>Iannini 2007</td>
<td>6</td>
<td>0.01</td>
</tr>
<tr>
<td>Jehl 1995</td>
<td>28</td>
<td>0.01</td>
</tr>
<tr>
<td>Maugein 2003</td>
<td>10</td>
<td>0.01</td>
</tr>
<tr>
<td>Maddox 2003</td>
<td>15</td>
<td>0.01</td>
</tr>
<tr>
<td>Moroney 2001</td>
<td>50</td>
<td>0.01</td>
</tr>
<tr>
<td>Munson 2007</td>
<td>8</td>
<td>0.01</td>
</tr>
<tr>
<td>Pallares 1995</td>
<td>55</td>
<td>0.01</td>
</tr>
<tr>
<td>Sangthawans 2003</td>
<td>7</td>
<td>0.01</td>
</tr>
<tr>
<td>Song 2004</td>
<td>18</td>
<td>0.01</td>
</tr>
<tr>
<td>Yigla 1995</td>
<td>3</td>
<td>0.01</td>
</tr>
<tr>
<td>Yu 2003</td>
<td>45</td>
<td>0.01</td>
</tr>
<tr>
<td>Turett 1999</td>
<td>8</td>
<td>0.01</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>262</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

### 8.1.4 Peritonitis

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Events</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capdevila 2001</td>
<td>5</td>
<td>0.01</td>
</tr>
<tr>
<td>Martinez 2002</td>
<td>7</td>
<td>0.01</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>5</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### 8.1.5 Endocarditis

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Events</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martinez 2002</td>
<td>7</td>
<td>0.01</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>7</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### 8.1.6 Bacteraemia

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Events</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Castillo 2000</td>
<td>2</td>
<td>0.01</td>
</tr>
<tr>
<td>Gomez 1995</td>
<td>6</td>
<td>0.01</td>
</tr>
<tr>
<td>Ho 2006</td>
<td>21</td>
<td>0.01</td>
</tr>
<tr>
<td>Kim 2002</td>
<td>18</td>
<td>0.01</td>
</tr>
<tr>
<td>Kumashi 2005</td>
<td>8</td>
<td>0.01</td>
</tr>
<tr>
<td>Maddox 2003</td>
<td>2</td>
<td>0.01</td>
</tr>
<tr>
<td>Maugine 2003</td>
<td>24</td>
<td>0.01</td>
</tr>
<tr>
<td>Turett 1999</td>
<td>8</td>
<td>0.01</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>415</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

### Figure 2.1 Forest Plot of Pooled ORs grouped by syndrome
<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Experimental Events</th>
<th>Control Events</th>
<th>Weight</th>
<th>Odds Ratio M-H, Fixed, 95% CI</th>
<th>Odds Ratio M-H, Fixed, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2.1.1 US</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ailani 2002</td>
<td>1</td>
<td>26</td>
<td>52</td>
<td>0.4%</td>
<td>0.65 [0.06, 6.61]</td>
</tr>
<tr>
<td>Arditi 1998</td>
<td>2</td>
<td>35</td>
<td>12</td>
<td>1.0%</td>
<td>0.68 [0.14, 3.17]</td>
</tr>
<tr>
<td>Castilho 2000</td>
<td>2</td>
<td>34</td>
<td>23</td>
<td>1.7%</td>
<td>0.37 [0.08, 1.64]</td>
</tr>
<tr>
<td>Feikin 2000</td>
<td>51</td>
<td>364</td>
<td>383</td>
<td>14.2%</td>
<td>1.31 [0.95, 1.79]</td>
</tr>
<tr>
<td>Fiore 2000</td>
<td>2</td>
<td>22</td>
<td>13</td>
<td>1.1%</td>
<td>0.57 [0.12, 2.73]</td>
</tr>
<tr>
<td>Kaplan 1998</td>
<td>4</td>
<td>221</td>
<td>15</td>
<td>1.2%</td>
<td>1.29 [0.42, 3.91]</td>
</tr>
<tr>
<td>Kumashiro 2005</td>
<td>8</td>
<td>44</td>
<td>11</td>
<td>1.4%</td>
<td>1.47 [0.55, 3.99]</td>
</tr>
<tr>
<td>Metlay 2000</td>
<td>10</td>
<td>44</td>
<td>32</td>
<td>2.6%</td>
<td>1.07 [0.48, 2.39]</td>
</tr>
<tr>
<td>Moroney 2001</td>
<td>10</td>
<td>44</td>
<td>32</td>
<td>2.6%</td>
<td>1.07 [0.48, 2.39]</td>
</tr>
<tr>
<td>Mufson 2007</td>
<td>15</td>
<td>98</td>
<td>12</td>
<td>1.8%</td>
<td>2.18 [0.98, 4.89]</td>
</tr>
<tr>
<td>Turett 1999</td>
<td>8</td>
<td>19</td>
<td>64</td>
<td>0.8%</td>
<td>3.84 [1.49, 9.92]</td>
</tr>
<tr>
<td>Winston 1999</td>
<td>6</td>
<td>65</td>
<td>31</td>
<td>1.7%</td>
<td>1.25 [0.50, 3.12]</td>
</tr>
<tr>
<td><strong>Subtotal (95% CI)</strong></td>
<td>1016</td>
<td>6307</td>
<td>30.5%</td>
<td>1.27 [1.02, 1.59]</td>
<td></td>
</tr>
<tr>
<td><strong>Total events</strong></td>
<td>119</td>
<td>631</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Heterogeneity:</strong></td>
<td>Chi² = 12.04, df = 11 (P = 0.03);</td>
<td>I² = 9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect:</td>
<td>Z = 2.17 (P = 0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2.1.2 Mexico</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gomez-Barreto 1999</td>
<td>1</td>
<td>11</td>
<td>4</td>
<td>0.5%</td>
<td>0.57 [0.06, 5.81]</td>
</tr>
<tr>
<td>Gomez-Barreto 2000</td>
<td>11</td>
<td>25</td>
<td>7</td>
<td>0.9%</td>
<td>1.91 [0.58, 6.23]</td>
</tr>
<tr>
<td><strong>Subtotal (95% CI)</strong></td>
<td>36</td>
<td>51</td>
<td>1.4%</td>
<td>1.45 [0.52, 4.00]</td>
<td></td>
</tr>
<tr>
<td><strong>Total events</strong></td>
<td>12</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Heterogeneity:</strong></td>
<td>Chi² = 0.82, df = 1 (P = 0.36);</td>
<td>I² = 0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect:</td>
<td>Z = 0.72 (P = 0.47)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2.1.3 Spain</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspa 2004</td>
<td>42</td>
<td>229</td>
<td>50</td>
<td>6.6%</td>
<td>1.61 [1.03, 2.52]</td>
</tr>
<tr>
<td>Capdevila 2001</td>
<td>3</td>
<td>4</td>
<td>17</td>
<td>0.2%</td>
<td>2.65 [0.25, 28.24]</td>
</tr>
<tr>
<td>Ewig 1999</td>
<td>8</td>
<td>57</td>
<td>3</td>
<td>0.6%</td>
<td>2.67 [0.67, 10.65]</td>
</tr>
<tr>
<td>Falco 2004</td>
<td>16</td>
<td>66</td>
<td>23</td>
<td>2.1%</td>
<td>2.20 [1.08, 4.48]</td>
</tr>
<tr>
<td>Gomez 1995</td>
<td>6</td>
<td>23</td>
<td>8</td>
<td>0.9%</td>
<td>1.76 [0.53, 5.86]</td>
</tr>
<tr>
<td>Martinez 2002</td>
<td>7</td>
<td>24</td>
<td>15</td>
<td>1.8%</td>
<td>0.66 [0.22, 1.96]</td>
</tr>
<tr>
<td>Pallares 1995</td>
<td>55</td>
<td>154</td>
<td>85</td>
<td>7.4%</td>
<td>1.79 [1.19, 2.70]</td>
</tr>
<tr>
<td>Pallares 2002</td>
<td>22</td>
<td>129</td>
<td>50</td>
<td>5.7%</td>
<td>1.03 [0.59, 1.78]</td>
</tr>
<tr>
<td><strong>Subtotal (95% CI)</strong></td>
<td>686</td>
<td>1420</td>
<td>25.4%</td>
<td>1.55 [1.23, 1.96]</td>
<td></td>
</tr>
<tr>
<td><strong>Total events</strong></td>
<td>159</td>
<td>251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Heterogeneity:</strong></td>
<td>Chi² = 6.76, df = 7 (P = 0.45);</td>
<td>I² = 0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect:</td>
<td>Z = 3.74 (P = 0.0002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 2.1.4 Asia

<table>
<thead>
<tr>
<th>Study</th>
<th>Total (95% CI)</th>
<th>Test for overall effect</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choi 1998</td>
<td>3 40 4.03 [0.40, 40.79]</td>
<td>Z = 4.51 (P &lt; 0.00001)</td>
<td></td>
</tr>
<tr>
<td>Ho 2006</td>
<td>21 135 0.89 [0.48, 1.67]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim 2002</td>
<td>18 77 0.75 [0.37, 1.55]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ma 2002</td>
<td>10 26 2.00 [0.58, 6.86]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sangthawan 2003</td>
<td>7 27 2.57 [0.67, 9.86]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Song 2004</td>
<td>18 105 1.16 [0.54, 2.49]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yu 2003</td>
<td>45 598 1.69 [1.13, 2.53]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>560 1008 1.32 [1.01, 1.73]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total events 122

Heterogeneity: Chi² = 7.68, df = 6 (P = 0.26); I² = 22%

Test for overall effect: Z = 2.03 (P = 0.04)

### 2.1.5 France, Italy UK

<table>
<thead>
<tr>
<th>Study</th>
<th>Total (95% CI)</th>
<th>Test for overall effect</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jehl 1995</td>
<td>28 263 0.72 [0.43, 1.20]</td>
<td>Z = 2.84 (P = 0.005)</td>
<td></td>
</tr>
<tr>
<td>Maddox 2003</td>
<td>5 91 0.80 [0.25, 2.54]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mauglein 2003</td>
<td>64 133 1.00 [0.61, 1.66]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pagliano 2007</td>
<td>1 54 5.89 [0.34, 102.88]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wasier 2005</td>
<td>9 32 1.65 [0.49, 5.54]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>546 573 0.91 [0.66, 1.25]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total events 107

Heterogeneity: Chi² = 3.58, df = 4 (P = 0.47); I² = 0%

Test for overall effect: Z = 0.59 (P = 0.55)

### 2.1.6 Canada

<table>
<thead>
<tr>
<th>Study</th>
<th>Total (95% CI)</th>
<th>Test for overall effect</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kellner 2002</td>
<td>1 42 0.46 [0.05, 4.70]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quach 2002</td>
<td>1 108 1.00 [0.10, 9.93]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>65 150 0.67 [0.13, 3.41]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total events 2

Heterogeneity: Chi² = 0.21, df = 1 (P = 0.64); I² = 0%

Test for overall effect: Z = 0.48 (P = 0.63)

### 2.1.7 Africa

<table>
<thead>
<tr>
<th>Study</th>
<th>Total (95% CI)</th>
<th>Test for overall effect</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friedland 1992</td>
<td>13 49 2.64 [0.99, 7.08]</td>
<td>Z = 2.03 (P = 0.039)</td>
<td></td>
</tr>
<tr>
<td>Molyneux 2002</td>
<td>22 143 2.08 [1.05, 4.12]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>75 192 2.25 [1.29, 3.94]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total events 35

Heterogeneity: Chi² = 0.15, df = 1 (P = 0.70); I² = 0%

Test for overall effect: Z = 2.84 (P = 0.005)

Total (95% CI) 2984 9701 100.0% 1.32 [1.17, 1.49]

Total events 556

Heterogeneity: Chi² = 42.25, df = 37 (P = 0.25); I² = 12%

Test for overall effect: Z = 4.51 (P < 0.00001)

Test for subgroup differences: Not applicable

---

**Figure 2.2 Geographical Grouping of ORs**
2.3.4 Adjusted ORs

Few studies reported adjusted ORs calculated from a statistical model. This was because antibiotic resistance was rarely a significant indicator of mortality in the univariate analysis. Table 2.4 presents the adjusted ORs for the impact of PNSP infection on mortality for the four studies which reported. Where studies reported ORs no confidence intervals were reported. Where the ORs are reported they are consistent and within the confidence interval of the pooled unadjusted OR found by the meta-analysis. The studies had all been adjusted for age, but only two had adjusted for underlying illness. None of the models found PNSP infection to be a significant indicator of mortality once the model had been adjusted for confounding variables.

Table 2.4 Adjusted OR

<table>
<thead>
<tr>
<th>Reference</th>
<th>OR</th>
<th>P Value</th>
<th>Variables Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Winston et al. 1999)</td>
<td>1.25</td>
<td>0.82</td>
<td>Age, Gender, Homelessness</td>
</tr>
<tr>
<td>(Yu et al. 2003)</td>
<td>1.42</td>
<td>0.19</td>
<td>Age, Critical Illness, Immunosupression, Chronic Disease</td>
</tr>
<tr>
<td>(Feikin 2003)</td>
<td>1.3</td>
<td>NS</td>
<td>Age</td>
</tr>
<tr>
<td>(Pallares et al. 1995)</td>
<td>1.3</td>
<td>0.84</td>
<td>Age</td>
</tr>
</tbody>
</table>

PNSP infections vs. PSP Infections
2.4 Discussion

This systematic review and meta-analysis examined published literature to analyse whether patients infected with an antibiotic resistant strain of *S. pneumoniae* had an increased risk of mortality compared with patients infected with antibiotic sensitive strains. One finding of the systematic search was that the majority of studies recorded whether the pneumococcal isolate was PNSP or PSP. Thus, this meta-analysis examined whether PNSP infections had an increased risk of mortality compared to PSP infections.

2.4.1 Impact of PNSP is syndrome dependent

The impact of PNSP infections varied dependent on the syndrome (pneumonia, bacteraemia and meningitis) assessed. Therefore, I will interpret the pooled ORs of each syndrome in turn.

1. Invasive pneumonia

The synthesis of published studies on mortality for patients with invasive pneumococcal pneumonia showed an association between PNSP pneumonia and an increased risk of mortality (OR 1.48, 95% CI 1.24-1.77). I also considered a subgroup of studies which grouped invasive pneumonia with bacteraemia; referred to as non meningeal IPnI. The combined OR for these studies also showed an association between PNSP IPnI and an increased risk of mortality (OR 1.27 CI 1.04-1.57). These results are consistent with the result of a previous meta-analysis on the effect of antibiotic resistant pneumococcal pneumonia (142).

The results from both sub-groups indicate an association between PNSP and increased risk of mortality, however it is not necessarily appropriate to interpret this as a causative relationship between PNSP infection and excess mortality. A number of studies in the meta-analysis did not appropriately control for confounding from age or severity of underlying illness or increased age within the study, which are in turn associated with increased severity of pneumonia. A common feature of the cohort and case control studies which considered pneumococcal pneumonia was that the patient populations examined ranged in age from young adults to the elderly, and the populations were not stratified by age (159,160,168,172,180,183,184,188,189,191,192). Previous research has shown that PNSP infection is associated with increased age as well as underlying medical conditions (193), so any apparent association between PNSP infection and increased mortality could be an association between age or underlying illness and increased risk of mortality.
In those studies where an OR was calculated using a regression model which allows the authors to control for confounding, PNSP infection was never found to have a significant negative impact on mortality once confounders such as underlying illness and age were controlled for (187,194). The results of the multivariate models indicate that any association between PNSP and increased risk of mortality is a function of the known associations between underlying disease and antibiotic resistance, these underlying factors increase the risk of mortality independent of the susceptibility status of the infecting strain.

If there is no significant causal association between PNSP infection and increased mortality, then it follows that discordant therapy is unlikely to have a negative impact. Most studies examining pneumococcal pneumonia reported that patients were not treated with penicillin. The most common reported treatment was cephalosporin (159,160,168,172,183,184,188,189,191,195,196). In those studies where patients were treated with penicillin, there was no increased risk of mortality and thus, no evidence for the negative impact of discordant therapy (34,188).

ii. Bacteraemia

The combined OR for studies of bacteraemia was 1.03 (95% CI 0.77, 1.37). This indicates that PNSP bacteraemia is not associated with a higher risk of mortality than patients with PSP bacteraemia. This is a reasonable interpretation as while all studies tested the bacteria for penicillin susceptibility the treatment of choice was a third generation cephalosporin or vancomycin. Therefore patients did not experience a delay in treatment. (189). Only one other study reported the treatment regime of the patients. Ho et al. reported that the majority of patients in their study were initially treated with penicillin (197). This study did not report any negative impact of treatment on mortality.

iii. Meningitis

The combined OR for all studies which examined patients with meningitis was 1.54 (95% CI 1.01-2.34). This indicates that PNSP meningitis infections are associated with an increased risk of mortality compared to PSP infections. Unlike the studies examining bacteraemia and pneumonia, the patient population included in the meningitis cohorts were all children, thus age was a not a confounding factor in these studies. If this association was a true indication of the negative impact of PNSP meningitis one would expect to see a negative impact of discordant therapy.

The recommended initial treatment in the United States, Europe for pneumococcal meningitis is third generation cephalosporin or vancomycin. In studies carried out in the US and Europe analysed in this meta-analysis, patients were treated with third generation cephalosporin or vancomycin. In
all these studies the authors reported the association between PNSP infection and mortality. The combined OR for these studies was 0.96 (95% CI 0.47-1.94) (152,161,198,186,179). This indicates that if patients are initially treated with appropriate therapy PNSP infection carries no excess risk.

When patients are not treated with concordant therapy the outcome is very poor. Studies carried out in Africa reported that patients with PSP and PNSP meningitis were treated with penicillin (162,176), the patients with PNSP meningitis had over twice the risk of mortality compared to patients with PSP meningitis (OR 2.25, 95% CI 1.29-3.94) (162,176).

2.4.2 PNSP meningitis places an excess burden of mortality in regions with routine penicillin use

Klugman et al. highlighted that penicillin and chloramphenicol remains widely used in the treatment of meningitis in developing countries (112). The possible clinical impact of an increased mortality due to PNSP meningitis is dependent on the prevalence of PNSP pneumococci in carriage and disease in a region. Currently national systematic surveillance systems of the prevalence and incidence of PNSP pneumococci are largely restricted to Europe and North America (199), though the SENTRY surveillance system was established in 1997 to monitor the prevalence of antimicrobial resistance in clinically relevant pathogens in the United States, Canada, Europe, Latin America, and the Asia-Pacific region (200). Such surveillance systems estimate the prevalence of antibiotic resistance (particularly to penicillin and erythromycin) in invasive disease isolates and indicate that PNSP is a global problem. PNSP has been shown to be of clinical relevance in studies carried out in Sub-Saharan Africa. In one such study 70% of isolates from respiratory infections in South Africa were intermediately resistant to penicillin (136) and in another study in rural Malawi 17% of cases of pneumococcal meningitis were intermediately resistant to penicillin (112). A survey of published literature found that the average prevalence of intermediate penicillin resistance from studies in Central Africa was 13% (201).

Although it was known for some time that meningitis was responsible for much child mortality worldwide, this has been quantified systematically in a recent study. O’Brien et al. calculated the burden of disease caused by S. pneumoniae in children younger than 5 years, and estimated that in 2000, 14.5 million episodes of serious pneumococcal disease occurred worldwide (203). The greatest number of cases occurred in Asia, while the greatest number of cases per head of population under the age of five occurs in Africa (3). An estimated 43,100 cases of pneumococcal meningitis occurred in the African region (as defined by WHO region) and the incidence rate is estimated as 38 cases per 100,000 children under 5 (203). Thus, given that pneumococcal meningitis is a serious public health problem in Africa, and the clinical impact of PNSP meningitis is so severe if
incorrectly treated there is an urgent need to survey the prevalence of PNSP at a national level in this region and to allow physicians resources to efficiently treat meningitis with appropriate concordant therapy.

2.4.3 Penicillin is still an effective treatment for many episodes of pneumococcal pneumonia and bacteraemia

All studies used in this analysis calculated the susceptibility of the infection to penicillin treatment using the CLSI guidelines published prior to 2008. Before January 2008 isolates susceptible to penicillin were defined as those with a MIC of $\leq 0.06 \mu g/ml$. As I stated in the introduction the effect a dosage of penicillin has on an infection depends on the site of treatment (65). If a drug is to be clinically effective then the concentration at the site of infection (blood stream, lungs or central nervous system) must exceed the MIC of the isolate 95% of the time. As penicillin penetrates poorly into the cerebrospinal fluid, the dosage of penicillin at the site of infection for meningitis will be much less than an infection in the lungs or blood stream. Thus a dosage of penicillin which may be suitable for treatment of a PNSP bacteraemia will be not suitable to treat PNSP meningitis. This explains the effects of treatment seen in this meta-analysis.

In January 2008 the CLSI guideline MICs were changed to reflect the true clinical impact of penicillin non-susceptibility, MICs for penicillin non-susceptibility in meningitis remained the same while the MICs for non-meningitis infections when treated intravenously were increased (18). The results of this meta-analysis are supportive of this change. Only studies examining PNSP meningitis infections reported a negative outcome as a result of discordant therapy, while patients infected with PNSP meningitis or bacteraemia showed no increased risk of mortality as a result of treatment with penicillin. On this basis these revised breakpoints penicillin administered intravenously is still an effective treatment for most invasive pneumonia and bacteraemia.

Mortality is one important but narrow definition of impact and cost, and due to a paucity of study I did not attempt to quantify the increased cost of treatment or length of stay of patients with PNSP infection. A consequence of the unnecessary use of third generation cephalosporin and vancomycin to treat pneumococcal pneumonia and bacteraemia is that an unnecessary economic cost is being imposed on health systems.

A caveat to the optimistic analysis above is the problem of truly penicillin resistant pneumococci. While in resource rich settings these can still be effectively treated with alternative therapy such a third generation cephalosporin and vancomycin, these alternative treatments may not be routinely available in resource poor settings. The incidence of pneumococcal pneumonia in children under 5 in
the year 2000 is estimated as 3,397 per 100,000 in Africa and 2,911 per 100,000 in South East Asia (202). An increase in the prevalence of penicillin resistant infections could result in an increased rate of mortality or greatly increased economic burden due to treatment provision.

2.5 Conclusion

At the outset the aim of this meta-analysis was to examine the impact of antibiotic resistance on the risk of mortality. The results showed that penicillin non-susceptibility increased the risk of mortality two fold when patients with non-susceptible meningitis were treated with penicillin.

Firstly this result demonstrates the importance of treatment guidelines for clinicians that take account of the effect of non-susceptibility on treatment. Such guidelines are influenced by the prevalence of resistance seen in patients and so are dependent on accurate surveillance of antibiotic resistance. In countries where penicillin treatment of meningitis is routine due to lack of awareness of the prevalence of PNSP, or lack of resources to treat routinely with more treatment, such as third generation cephalosporin or vancomycin, the risk of death increases two fold for a patient with PNSP meningitis compared to a patient with PSP meningitis.

Secondly these results warn of the impact of the loss of third generation cephalosporins as effective treatment of meningitis due to development of resistance to this antibiotic. They show that once effective treatment options are all but exhausted antimicrobial resistance will result in an increased risk of mortality.
3 Explaining heterogeneous prevalence of resistance among serotypes: Hypothesis and Mathematical Model

3.1 Introduction

In the previous chapter, I showed that PNSP infection can be clinically costly. In pneumococcal meningitis, the probability of death is doubled for patients with a PNSP infection inappropriately treated with penicillin. Although the impact of antibiotic non-susceptibility can be overcome by changing treatment regimen, this is not always possible and so patients may still be inappropriately treated, especially in resource poor settings. Thus, interventions to prevent the emergence and spread of antibiotic resistance are necessary and so understanding how antibiotic resistance evolves in a pneumococcal population is of direct public health importance. In this chapter I will shift focus onto the study of the basic mechanisms which regulate the spread of antibiotic resistance at the population level.

Broadly, the question of what determines the prevalence of resistance is separate from that of understanding how genetic mechanisms of resistance evolved in the first place. The prevalence of resistance is determined by a trade-off on the one hand between selective pressure exerted by human consumption of antibiotics, and on the other the biological cost of resistance which reduces the ability of resistant strains to spread. The details of this trade off are quantitative and specific to each case. We will therefore need a detailed mathematical model of resistance and antibiotic consumption in the pneumococcal population to help interpret trends.

3.1.1 Antibiotic consumption occurs independently of carriage

The majority of carriage and transmission of pneumococci occurs amongst children under the age of two. The majority of episodes of carriage do not lead to disease (204). Infants and children under two years of age have very high rates of antibiotic consumption (205) and these high rates of antibiotic consumption, combined with high rates of carriage, place a selective pressure on pneumococci to develop and maintain antibiotic resistance. Resistant strains of pneumococcus maintained in the carriage population can then occasionally cause invasive disease in all age groups, so prevalence of resistance in invasive disease is related to that found in carriage.

Because antibiotic consumption occurs independently of carriage, mathematical models of resistance in the pneumococcus will differ in an important respect from models of resistance in
obligate pathogens where treatment occurs in direct response to infection. For the pneumococcus antibiotic consumption and carriage should be modelled as uncoupled independent processes.

3.1.2 The importance of serotype

Previously published surveys of pneumococcal carriage have reported uneven distributions of antibiotic resistance among serotypes. Some serotypes are more likely to carry antibiotic resistance than others and this pattern is conserved among surveys from different countries and settings (90). Broadly 6B, 9V, 14, 19F, 23F, 6A and 19A are frequently associated with antibiotic resistance and serotypes 3, 18C, 15A and 1 are more rarely associated with antibiotic resistance (90).

This phenomenon has been illustrated particularly well in a recently published study of pneumococcal invasive disease in Spain, which compiled surveillance of IPD from 1979 to 2007 (193). This survey examined the prevalence of PNSP in invasive disease (which is the prevalence of isolates that were intermediately susceptible to penicillin and those which were resistant). During the 1990s the prevalence of PNSP was stable at 42% for all isolates. However serotypes 1, 3, 4, 5, 7F and 8 had no reported penicillin non-susceptibility, whereas serotypes 6B, 9V, 14, 19F and 23F were 90 -100% PNSP (193).

This simple but striking observation leads us to consider which aspects of serotype biology could explain these observed heterogeneities in resistance.

The three important characteristics of serotypes to consider are:

- Invasiveness,
- Serotype-specific age distribution of carriage and disease,
- Prevalence and duration of carriage.

i. Invasiveness

As discussed previously, certain serotypes are more likely to cause disease than others. The relative attack rates (relative invasiveness of a serotype) of pneumococcal serotypes were measured using national UK surveillance data on IPD and serotype-specific acquisition rates collected from a longitudinal study of carriage. This study found significant differences between attack rates. For example serotypes 4, 13, 7F, 9V and 18C were associated with 120 invasive disease cases /100,000 acquisitions and serotypes 23F, 6A, 19F, 16F, 6B and 15B/C were associated with 10 invasive disease cases /100,000 acquisitions (31). However, differences in a serotype’s invasiveness does not seem to be a plausible biological mechanism to explain differences in resistance since, as discussed above, there is no strong link between antibiotic consumption and carriage.
ii. **Age**

Traditionally serotypes thought to be particularly associated with carriage in children have been characterised as ‘childhood’ serotypes. Serotypes which are commonly isolated from the nasopharynx of children would be under greater antibiotic pressure than serotypes isolated more commonly from adults, as children consume more antibiotics. Therefore any study which looks for an association between serotypes and resistance should be stratified by age to avoid spurious associations. The impact of age will be explored in Chapter 4.

iii. **Serotype prevalence and duration**

Authors have frequently noted that the more prevalent serotypes have a higher prevalence of antibiotic resistance (193,56). This in turn is linked to the duration of time a serotype is carried in an individual; serotypes that are carried for a longer time are more likely to be present in the sampled population. Previous studies have consistently shown that different serotypes are carried for different amounts of time in the host (206,207,31,208). Although there is variation in serotype-specific duration of carriage estimated from different studies, certain serotypes are consistently characterised as having a short duration of carriage (such as serotype 4, (31,209,210)) or a long duration of carriage (such as serotype 6A (31,209,210)). Different serotypes are also present at different prevalences in the population (206,207,31,208).

Recently an interesting biological mechanism which explains differences in serotype duration has been discovered (18). In the study, the authors hypothesised that serotypes that had less complex capsular polysaccharide structures may be more successful at evading the immune system, and so persist longer in the host and at higher prevalence in the population. Weinberger et al. showed that serotypes with less complex polysaccharides were more heavily encapsulated and that heavily encapsulated serotypes were less susceptible to clearance by the immune system of the host (18). When the authors compared their findings to published data on serotype prevalence and carriage duration, heavily encapsulated serotypes were found to have longer durations of carriage and to be more prevalent in the population (18).
3.2 Hypothesis - distribution of antibiotic resistance amongst serotypes

Taking previously published knowledge of serotype characteristics, a hypothesis was developed to explain the distribution of antibiotic resistance among serotypes. Figure 3.1 illustrates schematically the effect of antibiotic resistance on two serotypes with two different durations of carriage (the period of time between acquisition and natural clearance by the immune system). The natural duration of carriage is thought to be the same in a resistant and sensitive strain of the same serotype in the absence of antibiotic consumption. When an individual consumes an antibiotic, the sensitive strain is cleared from the nasopharynx and has its duration of carriage cut short. The resistant strain is carried for its natural duration, irrespective of antibiotic use. As the resistant strain is carried longer than the sensitive strain, the resistant strain has more opportunity for transmission and thus becomes more prevalent as a constituent strain of the serotype. The fitness advantage of a resistant strain is thus the average difference in carriage duration between it and the sensitive strain.

From Figure 3.1, I hypothesise that the relative fitness advantage is greatest in serotypes with longer durations of carriage. Therefore a serotype with a long duration of carriage gains a larger advantage in being resistant than a serotype with a short duration of carriage. As a result, one would expect there to be a positive relationship between the duration of carriage of a serotype and the presence of antibiotic resistance.
As I discussed in the introductory chapter, antibiotic resistance is often associated with reduced fitness of the resistant strain compared to the sensitive strain (211). In communities where antibiotic use is absent or rare, the reduced fitness of the resistant strain means that they are outcompeted by sensitive strains. Conversely, for a resistant strain to become established in a community, the fitness advantage of antibiotic resistance must be greater than the fitness cost of resistance. On the other hand, because the genetic mechanism(s) of resistance are independent of serotype (59,212), the cost of resistance is plausibly assumed to be constant. This means that there should be a threshold fitness advantage (duration of carriage in this hypothesis) below which antibiotic resistance is not maintained in the population.

In species of bacteria where there is no difference in the duration of carriage between different strains, fitness advantage is the same in each strain. Previous mathematical models of antibiotic resistance examined these types of bacterial populations, where the fitness advantage of resistance is constant in all bacteria in carriage (14-17).
3.2.1 Aims of chapter

I will investigate whether the hypothesis outlined in the previous paragraph can adequately describe the distribution of antibiotic resistance among serotypes in the pneumococcal population. In this chapter I will explore the hypothesis using two mathematical models. I will:

- Explore the relationship between the duration of carriage of a serotype and the prevalence of resistance.
- Quantify the impact of antibiotic consumption on this relationship.
- Examine the interaction between the fitness (cost of resistance), the duration of carriage of a serotype and the rate of consumption of antibiotics.
3.3 Mathematical model of antibiotic resistance

In section 3.2, I outlined a hypothesis which provides an explanation for the heterogeneous distribution of antibiotic resistance among serotypes. In this section I explore that hypothesis using a mathematical model, and quantify what serotype characteristics promote the maintenance of antibiotic resistance. The aim is to make more precise the intuitive arguments outlined in 3.2.

3.3.1 Model Structure

I use two deterministic transmission models of pneumococcal carriage developed previously (213, 214). Both models are highly idealised representations of the strain structure within one serotype, with two strains of pneumococci, one susceptible to antibiotics and the other resistant. In both models pneumococci can be cleared from the host naturally and sensitive strains can also be cleared through exposure to antibiotics. These models describe transmission of strains within a single serotype. I then vary the natural duration of carriage of the serotype to simulate the characteristics of different serotypes. I do not consider the dynamics of multiple serotypes co-existing which, though relevant and interesting, would make the models too complex and intractable at this stage. Although this is a model of carriage rather than disease, for the sake of exposition I will use ‘infection’ as short hand to refer to carriage and acquisition of carriage for the remainder of the chapter.
Figure 3.2 Schematic illustration of two theoretical models.
$X$ indicates susceptible individuals, $I_s$, those infected with a sensitive strain, $I_r$, those infected with a resistant strain and $I_{sr}$ are individuals infected with both strains. Infection with one strain is indicated by an arrow, recovery is indicated with a dot dash arrow and consumption of antibiotics is indicated with a dashed arrow. In model A dotted arrows indicate re-infection. In model B there are two additional classes $I_{ss}$ and $I_{rr}$, where individuals are infected with two strains. Super-infection from a single to a double infection is indicated with a dotted arrow and from a dual infection to a dual infection with a small dashed arrow Reproduced from (84,85).

In previous work, Lipsitch et al. and Colijn et al. (213,214) analysed the theoretical foundations of multistrain epidemic models, and discovered that great care needs to be taken to avoid introducing spurious dynamics that arise from subtle choices in model structure. They suggested that a 'structurally neutral' model is defined by the condition that strain frequencies should not be under any selection when all parameters governing fitness are equal. Selection should only arise due to differences in fitness between strains. This condition leads to biologically counterintuitive model structures such as the $I_{ss}$ term in model B (see Figure 3.2). Both model A and B are examples of structurally neutral models. I consider both here as given current knowledge there are no grounds to distinguish between them.

1. **Model A**

The simpler of the two models, Model A, is illustrated in Figure 3.2 (A) and the equations of the model are shown as equation group 3-2. In this model individuals may be infected with a single strain or co-infected with both resistant and sensitive strains. As shown in Figure 3.2 (A) the basic model contains four types of individuals: $X$ (susceptible), $I_s$ (infection with a sensitive strain), $I_r$ (infection with a resistant strain) and $I_{sr}$ (co-infection).
Individuals are infected with a sensitive strain proportional to rate constant \( \beta_s \) and to the prevalence of individuals infected with a sensitive strain \((I_s)\) or individuals with dual infection \((I_{sr})\). The rate of infection with a resistant strain is proportional to its rate constant \( \beta_r \) and the prevalence of individuals infected with a resistant strain \((I_r)\) or individuals infected with two strains \((I_{sr})\). Individuals infected with both strains may transmit either strain on contact with a susceptible individual or individual infected with one strain. The force of infection \((\lambda_{s/r})\) is:

\[
\lambda_r = \beta_r (I_r + qI_{sr}) \\
\lambda_s = \beta_s (I_s + qI_{sr})
\]

In this model \( q = \frac{1}{2} \), meaning that dually infected individuals are equally infectious as the average infectiousness of singly infected hosts with the sensitive or resistant strain alone. A singly infected individual may become dually infected upon contact with the competing strain.

Re-infection with a second strain is less efficient than the initial infection. This is modelled through the parameter \( k \). Re-infection with a second strain occurs at a per capita rate of \( k \) times less likely where \( k \) is a number between \( \frac{1}{2} \) and 1. When \( k \) is within this range the model simulates partial resistance to super-infection, re-infection and co-infection play a role in the dynamics of the model. The impact of this on transmission is that individuals carrying a single strain \((I_s \text{ or } I_r)\) are partially protected against super infection with a second strain.

If a host infected with two strains \((I_{sr})\), is super infected with a “new infection” sensitive strain that host is then defined as a singly infected individual with a sensitive strain \((I_s)\). A resistant strain can also replace a sensitive strain in the same way. With probability \( c = \frac{1}{2} \), re-infection of a dually infected individual replaces the same strain, resulting in no change. Dually infected individuals have the same infectiousness as singly infected individuals so that \( q = c = \frac{1}{2} \). The choice of \( q = c = \frac{1}{2} \) means that the structural neutrality of the model is conserved (84).

An individual in class \( I_s, I_r \text{ or } I_{sr} \) returns to the susceptible class \( X \) at rate \( \mu \). This is the natural rate of clearance, defined as the inverse of the natural duration of carriage. Sensitive and resistant strains have the same natural duration of carriage in the absence of antibiotic consumption.

Antibiotic consumption is simulated in this model through the consumption rate \( \tau \). An individual infected with a sensitive strain may return to the class \( X \) and individuals in class \( I_{sr} \) may return to class \( I_r \) through antibiotic consumption. In the second case antibiotic consumption clears the
sensitive strains carried by the individual. Antibiotic consumption is assumed to be instantaneously effective therefore it can be considered as a drug mediated recovery rate.

The following set of equations describes the dynamics of $I_s$, $I_r$, and $I_{sr}$ with time:

3.2 Model A

\[
\frac{dI_s}{dt} = \lambda_s X + k c \lambda_s I_{sr} - k \lambda_r I_s - (\mu + \tau) I_s \\
\frac{dI_r}{dt} = \lambda_r X + \tau I_{sr} + k c \lambda_r I_{sr} - k \lambda_s I_r - \mu I_r \\
\frac{dI_{sr}}{dt} = k \lambda_r I_s + k \lambda_s I_r - (\mu + \tau) I_{sr} - k c \lambda_s I_{sr} - k c \lambda_r I_{sr} \\
X = 1 - I_s - I_r - I_{sr}
\]

ii. Model B

Model B is an unusual transmission model, as it allows both co-infection and co-transmission of resistant and sensitive strains (214,213). It builds upon model A in two ways, firstly re-infection may result in individuals having two "copies" of a single strain (states $I_s$ and $I_r$) and secondly this model allows dual transmission of two strains in a single transmission step.

As shown in Figure 3.2(B), Model B contains four types of individuals: $X$ (susceptible), $I_s$ (infected with a sensitive strain), $I_r$ (infected with a resistant strain) and $I_{sr}$ (dual infection). Building upon model A the model now has two additional states $I_s$ and $I_r$.

Infection with a sensitive strain is proportional to rate constant $\beta_s$ and the prevalence of individuals infected with a sensitive strain ($I_s$) or dually infected individuals ($I_{sr}$ and $I_{ss}$). The rate of infection with a resistant strain is proportional to its rate constant $\beta_r$ and the prevalence of individuals infected with resistant strain ($I_r$) or co-infected individuals ($I_{sr}$ and $I_{rr}$).

Extending Model A, Model B allows acquisition of two strains in a single transmission step. Susceptible individuals (the $X$ class) may enter the dual infected class ($I_{sr}$, $I_{ss}$ or $I_{rr}$) on contact with a dual infected individual (those who are $I_{sr}$, $I_{ss}$ or $I_{rr}$). In this case transmission of a single strain can occur with probability $\rho_{single}$. The transmission of a sensitive strain from a dually infected individual occurs with probability $\rho_s$. The equations for acquisition of carriage are shown in equation set 3-4. Those dual infected individuals are $q$ times as infectious with each of their strains. When $q = \frac{1}{2}$ as both Model A and Model B a dual infected state is as infectious as a singly infected state. The set of equations describing the dynamics of the model are presented below:
As discussed in the initial outline of the hypothesis (section 3.2), antibiotic resistance is thought to carry an associated fitness cost. Models A and B can incorporate a fitness cost at the initial transmission stage; in these models $\beta$ (the probability of infection on contact by the number of contacts) of the resistant strain is less than that of the sensitive strain. This makes it harder for a resistant strain to transmit in the population relative to the sensitive strain.

In model B there is a second method of incorporating a fitness cost in a dually infected person the sensitive strain is more likely to transmit than the resistant strain. When a dually infected individual ($I_{sr}$) is in contact with a susceptible individual ($X$) and a single infection is transmitted (with probability $\rho_{\text{single}}$), a sensitive strain is transmitted with probability $\rho_s$ and resistant strain is
transmitted otherwise. In both models it is assumed that the cost of resistance does not change with the serotype, therefore the cost of resistance is independent of the duration of carriage.

3.3.2 Parameters and simulations

Previous work with this model investigated the parameter space which allows the coexistence of susceptible and resistant strains of pneumococci (85). Within these restrictions I defined a set of plausible parameters. Where a parameter was available from published literature I used this in the model; if this was not possible I simulated the model for a range of values. The parameters for this model are shown in Table 3.2.

The range of carriage duration was informed by previous published work. As carriage duration has been shown to vary for different serotypes, I ran simulations for a range of carriage duration values estimated from a carriage study of children in Kenya (described in more detail in Chapter 4, Table 4.1). From this study the duration of carriage of a serotype ranged from 9 to 52 days.

Previous research using this model has examined a fixed duration of carriage and consequently fixed transmission parameters (213). An essential part of the research presented here is to examine the impact of variable rates of carriage, therefore transmission parameters varied with the duration of carriage.

Previous research has shown that prevalence of pneumococcal infection (asymptomatic carriage) in children under three years of age varies considerably in different communities (205,21,215). For the simulations presented in this analysis, I chose to simulate transmission, and thus a value of $\beta_s$ was chosen so that the prevalence of infection (proportion of individuals in any infected state) was maintained at 26%. This is the prevalence of infection when individuals in the population consume antibiotics at a rate of 1/3 per person per month. These prevalence and antibiotic consumption estimates are taken from a community in the United States (205).

The relationship between transmission and duration of any infection can be expressed though $R_0$ in a community with no antibiotic consumption, where $R_0$ is the number of secondary infections, from a primary infection in a fully susceptible population and $D$ is the duration of infection.

3.5 Relationship between $\beta$, $R_0$ and duration of infection

$$\beta_s = \frac{R_0}{D}$$

$R_0$ was maintained at a constant value of 1.5, which maintained a prevalence of infection of 33% when antibiotic consumption was zero. The values of $\beta_s$ varied from 4.22 to 1.01 per person per year.
for corresponding duration of infection of 2 to 56 days, in order to maintain a constant value of $R_0$.

When the antibiotic consumption rate was $1/3$ per person per month, the prevalence of infection (carriage) varied from 26%-0.16%, for these parameters.

I used antibiotic consumption rates estimated from studies of children in day care. This study found that children under three years of age consumed antibiotics (specifically penicillin) at an average rate of $1/3$ per person per month, and children under five consumed penicillin at an average rate of $1/12$ per person per month (205). These were taken as the upper and lower bounds of consumption in this analysis.

As there is no conclusive evidence of values of the transmission cost of antibiotic resistance in carriage, I simulated a range of parameter values, though I assume that the parameter value for the cost of resistance is not related to the duration of carriage. In Model A and Model B, the transmission parameter of the resistant strain ($\beta_r$) is proportional to the transmission parameter of the sensitive strain $\beta_s$, so that:

3.6 Relationship between $\beta_r$ and $\beta_s$

$$\beta_r = f \beta_s$$

where the parameter $f$ is varied between 0.5 and 1.

In Model B the default assumption is that the cost of resistance is comparable at both single and double transmission, so that:

3.7 Model B cost of resistance

$$\frac{\beta_r}{\beta_s} = \frac{1 - p_i}{p_i}$$

Table 3.1 Cost of resistance relationship between parameters in model A and model B

<table>
<thead>
<tr>
<th>$f$</th>
<th>$p_s$</th>
<th>$\frac{\beta_r}{\beta_s} = \frac{1 - p_i}{p_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.660</td>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
<td>0.600</td>
<td>0.6</td>
</tr>
<tr>
<td>0.7</td>
<td>0.588</td>
<td>0.7</td>
</tr>
<tr>
<td>0.8</td>
<td>0.555</td>
<td>0.8</td>
</tr>
<tr>
<td>0.9</td>
<td>0.525</td>
<td>0.9</td>
</tr>
</tbody>
</table>
In Model B, $\rho_s$ the probability of infection with a sensitive strain on contact with an $I_s$ individual was varied between 0.52 and 0.6 with corresponding values for the $f$ parameter presented in Table 3.2. As part of the sensitivity analysis I varied $\rho_s$ between 0.5 and 0.6 while keeping $f$ constant and then varied $f$ between 0.7 and 1 while keeping $\rho_s$ constant.

I constructed the model using differential equations in Berkley Madonna. I ran the model for 50 simulated years, with a timestep of 1 day. In each simulation one resistant and one sensitive individual were added to a population of fully susceptible individuals.
Table 3.2 Parameter Table for Model A and Model B

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value/Range</th>
<th>Units</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_r$</td>
<td>$f \beta_s$</td>
<td>Per person per year</td>
<td>Transmission parameter for the resistant strain, assumes resistant strain is never more transmissible than the sensitive strain</td>
<td></td>
</tr>
<tr>
<td>$f$</td>
<td>0.75-1</td>
<td>Proportion</td>
<td>Fitness proportional cost. assumes resistant strain is never more transmissible than the sensitive strain</td>
<td>Computed</td>
</tr>
<tr>
<td>$\tau$</td>
<td>$1/12$-$1/3$</td>
<td>Per person per month</td>
<td>Drug mediated recovery rate</td>
<td>(30)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$\frac{1}{\text{Duration of carriage}}$</td>
<td>Per person per month</td>
<td>Natural recovery rate</td>
<td></td>
</tr>
<tr>
<td>Duration of carriage</td>
<td>9-52</td>
<td>Days</td>
<td>Number of days between acquisition and natural clearance</td>
<td>(208)</td>
</tr>
<tr>
<td>$k$</td>
<td>$\frac{1}{2}$</td>
<td>Per capita rate</td>
<td>1-partial immunity</td>
<td>Assumption</td>
</tr>
<tr>
<td>$c$</td>
<td>$\frac{1}{2}$</td>
<td>Proportion</td>
<td>fraction of dual infected returning to $I_S$ (or $I_R$) upon re-infection with $S$ (or $R$)</td>
<td>Necessary for structurally neutral model in simple model when strains are indistinguishable (84).</td>
</tr>
<tr>
<td>$\rho_{\text{single}}$</td>
<td>$\frac{1}{2}$</td>
<td>Proportion</td>
<td>Probability that a single (rather than dual) infection is transmitted</td>
<td>Assumption</td>
</tr>
<tr>
<td>$\rho_S$</td>
<td>$\frac{1}{2}$-$1$</td>
<td>Proportion</td>
<td>Probability that $S$ is transmitted when a dually infected individual transmits single infection</td>
<td>Assumption</td>
</tr>
<tr>
<td>$q$</td>
<td>$\frac{1}{2}$</td>
<td>Proportion</td>
<td>Relative infectiousness with each strain for dually infected ($I_{sr}$)</td>
<td>Necessary for structurally neutral model in simple model when strains are indistinguishable (84).</td>
</tr>
</tbody>
</table>
3.4 Results

3.4.1 Results from Model A

This section presents results from Model A, a transmission model of pneumococcal asymptomatic carriage, referred to as infection in the description of the model. The effect of antibiotic consumption, serotype specific carriage duration and cost of resistance are investigated and presented in the following sections. Unless otherwise indicated, the parameters in the simulations presented here are as described in Table 3.1 and Table 3.2.

i. Relationship between antibiotic consumption and treatment

In Model A, increasing the rate of antibiotic consumption increases the prevalence of antibiotic resistance in the model population. Figure 3.3 shows the proportion resistant described by equation 3-8 plotted against duration of carriage for three antibiotic consumption rates: low (1/12 per person per month), medium (1/6 per person per month) and high (1/3 per person per month).

3-8 Proportion resistant

\[ Prop(R) = \frac{I_r + (I_{r+2}/2)}{(I_r + I_s + I_{sr})} \]

At the low rate of antibiotic consumption the prevalence of resistance is zero; at medium and high consumption rates the prevalence of resistance can reach 100% for some durations of carriage. Increasing the rate of antibiotic consumption increases the prevalence of resistance; a serotype with a carriage duration of 35 days will have a prevalence of resistance of 100% if antibiotic consumption is at a rate of 1/3 per month, while if consumption is at a rate of 1/6 per person per month prevalence of resistance is 50%.

ii. Relationship between serotype carriage duration and prevalence of resistance

Shown in Figure 3.3, at the medium and high antibiotic consumption rates (1/6 per person per month and 1/3 per person per month) the relationship between duration of carriage and prevalence of resistance is positive. For the lowest rate of antibiotic consumption (1/12 per person per month), antibiotic resistance is absent at all carriage durations.

Increasing the rate of antibiotic consumption reduces the duration of carriage under which antibiotic resistance is zero. At the rate of antibiotic consumption documented in children under three (1/3 per person per month), serotypes with a duration of carriage of under 20 days do not maintain resistance. At a lower rate of antibiotic consumption (1/6 per person per month) this threshold of carriage duration is 32 days.
iii. Cost of resistance

As illustrated in Figure 3.3 there is a positive relationship between the duration of carriage and the prevalence of antibiotic resistance. I also showed that the prevalence of antibiotic resistance at each point of duration of carriage increases as the rate of antibiotic consumption increases. I then explored what effect the cost of resistance has on the prevalence of resistance within this model. Figure 3.4 shows the relationship between $\frac{\beta_r}{\beta_z}$, and the prevalence of resistance. This figure shows that simulations where the rate of antibiotic consumption is varied from 1/12 per person per month to 1/3 per person per month, while the duration of carriage is constant at 31 days.
Figure 3.4 Cost of resistance: rates of antibiotic consumption.
This figure shows the relative fitness of the resistant strain (where 0 is unable to transmit and 1 is as fit as the sensitive strain) plotted against the proportion of resistant bacteria. In these simulations the duration of carriage was 31 days, the re-infection probability $k=0.5$ and the simulation ran for 150 years. The figure shows rates of antibiotic consumption of $1/3$ per person per month (solid line), $1/6$ per person per month (dotted line) and $1/12$ per person per month (dashed line).

Figure 3.5 Cost of resistance: duration of carriage.
This figure shows the relative fitness of the resistant strain plotted against the proportion of resistant pneumococci. In these simulations the rate of antibiotic consumption was $1/3$ per person per month, $k=0.5$ and the simulation ran for 150 years. This figure shows duration of carriage of 42 days (solid line), 31 days (dotted line) and 9 days (dashed line).
These simulations illustrate that when $\frac{\beta_r}{\beta_s} < 0.8$, no resistant strains remain in the model population when the rate of antibiotic consumption is 1/3 per person per month and under. At low rates of antibiotic consumption (1/12 per person per month) when $\frac{\beta_r}{\beta_s} < 0.9$, the resistant strain does not persist in the population. I then examined the interaction between the duration of carriage and fitness where the rate of antibiotic consumption was constant at 1/3 per person per month. Figure 3.5 shows the results of simulations where the duration of carriage was 9, 31 and 42 days. This figure shows that resistant strains are present at high prevalence in the population when $\frac{\beta_r}{\beta_s} \geq 0.5$, when the duration of carriage is 42 days and $\frac{\beta_r}{\beta_s} \geq 0.5$. If the duration of carriage is 9 days resistance does not persist if $\frac{\beta_r}{\beta_s} < 0.9$.

iv. D$_5$: cost of resistance and impact of antibiotic consumption

As shown in Figure 3.3, Figure 3.4 and Figure 3.5, there is a consistent relationship between the rate of antibiotic consumption, the cost of resistance, duration of carriage and prevalence of antibiotic resistance. To further quantify this relationship I estimated D$_5$, which is the lowest duration of carriage (in number of days) of a serotype where at least 5% of pneumococci are resistant to antibiotics. The estimates of D$_5$ for different costs of resistance and rates of antibiotic consumption are presented in Table 3.3.

<table>
<thead>
<tr>
<th>Rate of Antibiotic consumption</th>
<th>1/3 per person per month</th>
<th>1/6 per person per month</th>
<th>1/12 per person per month</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{\beta_r}{\beta_s}$</td>
<td>Model A</td>
<td>Model B</td>
<td>Model A</td>
</tr>
<tr>
<td>0.9</td>
<td>9</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>0.8</td>
<td>22</td>
<td>22</td>
<td>33</td>
</tr>
<tr>
<td>0.7</td>
<td>38</td>
<td>36</td>
<td>&gt; 52</td>
</tr>
<tr>
<td>0.6</td>
<td>&gt; 52</td>
<td>&gt; 52</td>
<td>&gt; 52</td>
</tr>
<tr>
<td>0.5</td>
<td>&gt; 52</td>
<td>&gt; 52</td>
<td>&gt; 52</td>
</tr>
</tbody>
</table>

The lower bound for duration of carriage examined was 9 days and the highest was 52 days as these are realistic durations estimated from carriage. Therefore in Table 3.3, when D$_5$ > 52, resistance was not seen at levels above 5% for any duration of carriage up to 52 days. For example when the cost
of resistance was 0.5 in Model A and Model B resistance was below 5% for all levels of antibiotic consumption.

3.4.2 Results from Model B
In this section I present the results from the more complex Model B, which allows dual transmission and multiple co-infected and super-infected states. Unless otherwise indicated I assume that the fitness of a strain transmitting from a singly infected individual is the same as transmission from a co-infected individual. This means that dually infected individuals are not any more infectious than singly infected individuals and is true for resistant and sensitive strains.

i. Relationship between serotype carriage duration and prevalence of resistance

The relationship between the duration of carriage of a serotype and the prevalence of antibiotic resistance in the Model B population is presented in Figure 3.6. This figure also presents the impact of antibiotic consumption on this relationship.

In this analysis the fitness of the sensitive strain relative to the resistant is described by the equation:

$$\frac{\beta_r}{\beta_s} = \frac{1-\rho_s}{\rho_r} = 0.8$$

This corresponds to parameters $f = 0.8$, and $\rho_s = 0.56$.

The results from this analysis are very similar to those from Model A with $f=0.8$. At the medium and high antibiotic consumption rates (1/6 per person per month and 1/3 per person per month) the relationship between duration of carriage and prevalence of resistance is positive, and for the lowest rate of antibiotic consumption (1/12 per person per month) antibiotic resistance is absent at all carriage durations.

Increasing the rate of antibiotic consumption increases the prevalence of resistance at each carriage duration; the results predict that a serotype with a carriage duration of 35 days will have a prevalence of resistance of 100% if antibiotic consumption is at a rate of 1/3 per person per month. If consumption is at a rate of 1/6 per person per month the prevalence of resistance is 59%.

Increasing the rate of antibiotic consumption also reduces the duration of carriage under which antibiotic resistance is zero. At the rate of antibiotic consumption documented in children under three (1/3 per person per month), serotypes with a duration of carriage of under 20 days do not maintain resistance. At a lower rate of antibiotic consumption (1/6 per person per month) this threshold of carriage duration is 31 days.
Duration of carriage in days is plotted against the proportion of pneumococci that are resistant to antibiotics. The proportion resistant is calculated as:

\[ \text{Prop}(R) = I_r + (I_{sr}/2) + I_{sz} + I_{rr} \] (I_r + I_{sr} + I_s + I_{sz} + I_{rr})

This is shown for a population with low antibiotic use (1/12 per person per month) where resistance is not present in the population (dashed line), a usage rate of 1/6 per person per month (dotted line) and 1/3 per person per month (solid line). \( k=0.5 \) and the simulation ran for 150 years.

ii. **Cost of resistance**

As illustrated in Figure 3.6 there is a positive relationship between the duration of carriage and the prevalence of antibiotic resistance. The prevalence of antibiotic resistance at each duration of carriage increases as the rate of antibiotic consumption increases. I then explored what effect the cost of resistance has on the prevalence of resistance within this model.

Figure 3.7 shows the relationship between the relative fitness of the sensitive strain and the prevalence of resistance. In this analysis, I assumed that the fitness cost of antibiotic resistance would affect the resistant strain at transmission from a singly infected or co-infected individual. These simulations illustrate that when a resistant strain transmits less than half as well as the sensitive strain no resistant strains remain in the model population when the rate of antibiotic consumption is 1/3 per person per month and under. At low rates of antibiotic consumption (1/12 per person per month), a resistant strain must be almost as fit as the sensitive strain in order to persist in the population.
Figure 3.7 Model B, cost of resistance: effect of treatment.
This figure shows the relative fitness of the resistant strain plotted against the proportion of resistant pneumococci. In these simulations the duration of carriage was 31 days, the re-infection probability $k=0.5$ and the simulation ran for 150 years. This figure shows duration of carriage of $1/3$ per person per month (solid line), $1/6$ per person per month (dotted line) and $1/12$ per person per month (dashed line).

Figure 3.8 Model B, cost of resistance: duration of carriage.
This figure shows the relative fitness of the resistant strain plotted against the proportion of resistant pneumococci. In these simulations the rate of antibiotic consumption was $1/3$ per person per month, the re-infection probability $k=0.5$ and the simulation ran for 150 years. This figure shows duration of carriage of 42 days (solid line), 31 days (dotted line) and 9 days (dashed line).
Figure 3.8 shows simulations where the duration of carriage is varied from 9 to 42 days, while the rate of antibiotic consumption is constant at 1/3 per person per month. Like the results from Model A, when the duration of carriage of the pneumococcus is 9 days the resistant strain must be as fit as the resistant strain to persist in the population. As the duration of carriage increases the cost of resistance that a resistant strain can carry while persisting in the population increases. The lower bound of cost for the resistant strain is 0.5. When the cost of resistance is greater, the resistant strain cannot persist for pneumococci carried for 42 days.

iii.  $D_5$: cost of resistance and impact of antibiotic consumption

As I showed in the results from Model A, there is clearly a consistent relationship between the rate of antibiotic consumption, the cost of resistance, duration of carriage and prevalence of antibiotic resistance. To further quantify this relationship I estimated $D_5$ for Model B. The estimates of $D_5$ for different costs of resistance and rates of antibiotic consumption are presented in Table 3.3. When the cost of resistance is consistent with Model A the values of $D_5$ are consistent (within two days) within Model A.
3.5 Exploration of key model assumptions and sensitivity analysis

Both models are an extremely simplified simulation of the complex process of transmission of pneumococci in a group of young children. Each model contains some key assumptions that I will discuss in this section.

i. Strain neutrality

Models A and B are ‘structurally neutral models’, meaning that in the absence of antibiotic consumption the resistant and sensitive strains are equally fit. In Model A this is achieved by keeping parameters $c$ and $q$ at 0.5, while Model B is a true ‘structurally neutral’ model. For this reason I chose to keep parameters $c$ and $q$ constant at 0.5 throughout all the analysis.

ii. Strain-specific immunity

In Models A and B strain specific immunity is incorporated through the parameter $k$. In analysis of Models A and B, to examine parameters under which the coexistence of resistant and sensitive strains was possible, the authors showed that strain-specific immunity was an important parameter governing co-existence (85). In Model A, an individual infected with strain $S$ can be re-infected with strain $R$ and become co-infected in state $I_{sr}$, therefore the transition term of $I_r$ to $I_s$ is:

Equation 3-9 Transition term $I_r$ to $I_s$ in Model A

$$k\beta_r(I_r + qI_{sr})I_s$$

In Model A it is assumed that an individual infected with both strains may be infected again with one of those strains. When this happens one strain will be ‘knocked out’ and the host will become singly infected, e.g. an $I_{sr}$ individual may become $I_s$ on infection with an $S$ strain. This occurs at the same rate as secondary infection, with half of all infections being undetectable as the infecting strain will be the same as that ‘knocked out’. In a second scenario, an individual infected with strain $S$ are less likely to become infected with an $S$ strain than an $R$ strain (and the converse is true). In this case ‘knocking out’ a strain becomes less likely and the transition term from $I_{sr}$ to $I_s$ would be:

Equation 3-10 Transition term $I_{sr}$ to $I_s$

$$ck_0\beta_r(I_s + qI_{sr})I_{sr}$$

where $k_0 < k$ and $k_0 = f_k$
In previous analyses the authors showed that decreasing ‘knock-outs’ by reducing the value of $k_0$, increases the probability of co-existence of sensitive and resistant strains. I performed the same analysis but examine whether decreasing $k_0$, changes the value of $D_5$.

Table 3.4 $D_5$ in days for changing values of $k_0$

<table>
<thead>
<tr>
<th>$f$</th>
<th>Model A</th>
<th>Model B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>0.75</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

Where $t = k_0/k$

\[ \frac{\beta_r}{\beta_s} = \frac{1-\rho_s}{\rho_s} = 0.8 \]

Previous work has shown that decreasing ‘knock out’ may increase the probability of co-existence (85) but as Table 3.4 shows, does not greatly change the duration of carriage above which resistant strains persist in the population.

**iii. Cost of resistance in Model B**

In Model B, the cost of resistance is applied to a resistant strain in two ways, firstly (like Model A) the transmission rate constant of the resistant strain is less than that of the sensitive strain so that:

\[ \frac{\beta_r}{\beta_s} < 1 \]

The cost of resistance is also applied on transmission from a co-infected individual. When one strain is transmitted from a co-infected individual a sensitive strain is transmitted with probability $\rho_s$. When $\rho_s > 0.5$, a sensitive strain is more likely to be transmitted from a co-infected individual than a resistant strain.

In the simulations presented in this chapter I have assumed that both forms of resistance are applied equally at each stage. This assumes that within-host competition between strains is equal to the between-host competition. Figure 3.9, shows that when the duration of carriage is at intermediate levels (21 days in Figure 3.9), changing the value of $\rho_s$ changes the persistence of resistance within the strain. As the value of $\rho_s$ increases the proportion of resistant strains decreases; under the parameters presented in Figure 3.9 when $\rho_s > 0.56$, and therefore $\frac{1-\rho_s}{\rho_s} > 0.7$, no resistant strains are present.
Figure 3.9 $\rho_s$ sensitivity analysis.
Value of $\rho_s$ is plotted against the proportion resistant strains for $\frac{\rho_r}{\rho_s} = 0.8$ and duration of carriage of 9 days (dashed line), 21 days (dotted line) and 42 days (solid line). In each simulation antibiotic consumption is 1/3 per person per month.

Table 3.5 $D_s$ for changing values of $\frac{(1 - \rho_s)}{\rho_s}$ relative to $\frac{\rho_r}{\rho_s} = 0.8$

<table>
<thead>
<tr>
<th>$g$</th>
<th>Antibiotic consumption per person per month</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/3$</td>
<td>$1/6$</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>35</td>
</tr>
<tr>
<td>0.8</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>$&gt;52$</td>
</tr>
<tr>
<td>0.7</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>$&gt;52$</td>
</tr>
<tr>
<td>0.6</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>$&gt;52$</td>
</tr>
<tr>
<td>0.5</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>$&gt;52$</td>
</tr>
</tbody>
</table>

$$g = \frac{(1 - \rho_s)}{\rho_s}$$
To further quantify the impact of varying $\rho_s$, while $\frac{\beta_s}{\beta_z}$ is constant, I calculated $D_s$, when $(1 - \rho_s)/\rho_s \leq \frac{\beta_z}{\beta_s}$, the results of which are presented in Table 3.5. When $(1 - \rho_s)/\rho_s$ is twice as small as $\frac{\beta_z}{\beta_s}$, the value of $D_5$ is increased by two days in simulations where the population has a high rate of antibiotic consumption (1/3 per person per month). At the lower rate of antibiotic consumption (1/6 per person per month) when $(1 - \rho_s)/\rho_s$ is smaller than $\frac{\beta_z}{\beta_s}$, $D_5$ increases from 35 days to being greater than 52 days.

iv. Transmission from a co-infected individual

In Model B the parameter $\rho_{\text{single}}$ is the probability that on contact with a co-infected or super-infected individual ($I_{sr}$, $I_{rr}$, or $I_{ss}$) there is transmission of one strain. All simulations presented in this analysis have $\rho_{\text{single}} = 0.5$, which is the symmetric form of the model (i.e. 50% of the time a single strain is transmitted from a dual infected individual). When using this model to explore the coexistence of sensitive and resistant pneumococcal strains Colijn et al. found that decreasing $\rho_{\text{single}}$, thus increasing the probability of dual transmission, increased the level of stable coexistence of sensitive and resistant strains (85).

Increasing co-existence could allow resistant strains to persist at low prevalence but at durations of carriage lower than expected at symmetry. To investigate this I varied $\rho_{\text{single}}$ while keeping the relative fitness of the sensitive strain and antibiotic consumption constant at 0.8 and 1/3 per person per month respectively. As $\rho_{\text{single}}$ increases (decreasing the probability that a single strain will be transmitted), the value of $D_5$ increases from 15 days when $\rho_{\text{single}}$ is 0.1 to 25 days when $\rho_{\text{single}}$ is 0.9.
3.6 Conclusions

In section 3.2, I outlined a hypothesis that provides an explanation of the uneven distribution of the prevalence of antibiotic resistance among pneumococcal serotypes. Using two deterministic models of pneumococcal transmission, I examined the relationship between carriage duration and the prevalence of resistance within a serotype.

i. There is a positive relationship between duration of carriage and prevalence of resistance

In the introduction to this chapter I described research which had consistently shown that pneumococcal serotypes differ in their duration of carriage [21,26,31]. I then outlined the hypothesis in section 3.2, that the relative fitness advantage of an antibiotic resistant strain is greatest in serotypes with long durations of carriage. Thus, there should be a positive relationship between the duration of carriage and the prevalence of antibiotic resistance. When I used two mathematical models to quantify this relationship, both models consistently predicted that a serotype with long duration of carriage was more likely to be antibiotic resistant.

ii. There is a threshold duration of carriage under which antibiotic resistance does not persist

The fitness cost associated with antibiotic resistance arises from alterations in the pneumococcus associated with mechanisms of resistance. As these mechanisms are serotype independent the cost of resistance is assumed to be constant. Therefore, as I outlined in section 3.2, there should be a threshold duration of carriage under which the cost of resistance outweighs the benefit and so no antibiotic resistance persists. This threshold was a consistent feature of Models A and B and was robust to sensitivity analysis.

iii. The threshold of carriage is consistent for a given rate of antibiotic consumption and cost of resistance

The threshold duration of carriage ($D_5$ in the analysis) was very sensitive to the parameter values used for the rate of antibiotic consumption and the cost of resistance in both models. Broadly $D_5$ decreased as antibiotic consumption increased, and, increased as the cost of resistance increased. The value of $D_5$ was consistent in both models for a given rate of antibiotic consumption and fitness cost. This relationship can be used to inform data analysis as I will describe in the next chapter.
4 Analysis of pneumococcal populations from invasive disease and carriage

4.1 Introduction

The mathematical model described in Chapter 3 allowed me to explore the hypothesis that there is a positive relationship between the prevalence of resistance in a serotype and the duration of carriage of that serotype. The relationship between duration of carriage and prevalence of resistance can be described approximately by a sigmoid curve, with a threshold of carriage duration under which resistant pneumococci do not persist in the model population. The persistence of resistant pneumococcal strains at intermediate carriage durations is determined by the level of treatment and the fitness cost of resistance. In this chapter I describe a test of the hypothesis explored by the mathematical model using a collection of *S. pneumoniae* isolates from invasive disease.

4.2 Data sets

4.2.1 Invasive Disease Data

I analysed data collected as part of the Active Bacterial Core Surveillance (ABCS) which is part of the United States Centers for Disease Control and Prevention’s Emerging Infections Program that has conducted active, population and laboratory based surveillance of invasive pneumococcal infection (defined as isolation of the bacteria from a usually sterile site such as blood, cerebrospinal fluid (CSF) or pleural fluid) since 1995.

Cases of pneumococcal invasive disease are recorded at least monthly through active communication with all microbiology laboratories which serve acute care hospitals. Using a standardized case reporting form information is recorded on patient demographic characteristics, clinical syndrome and geographic location to the county level (216).

Isolates are characterized at a reference laboratory through antimicrobial susceptibility testing using CLSI guidelines (217) and serotyping (50). The dataset used in this study has information on demographic and geographic characteristics, antimicrobial susceptibility and serotype data. I extracted the frequency of each serotype, the distribution of susceptibility to penicillin and erythromycin in that serotype, the age of the patient from whom it was isolated and the year of isolation.

I have assumed that the consumption of oral antibiotics is the primary driver of antibiotic resistance in pneumococci, thus I have interpreted the MICs of each isolate in line with CLSI guidelines for oral
antibiotics. MIC cut offs for susceptible, intermediate and resistant isolates are presented in Table 4.1.

**Table 4.1 Former and current Clinical and Laboratory Standards Institute Penicillin and erythromycin minimum inhibitory concentration (MIC) breakpoints**

<table>
<thead>
<tr>
<th>MIC $\mu g/ml$, by susceptibility category</th>
<th>Period, syndrome and route of administration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Susceptible</td>
</tr>
<tr>
<td><strong>Penicillin</strong></td>
<td></td>
</tr>
<tr>
<td>Before January 2008</td>
<td></td>
</tr>
<tr>
<td>All syndromes and routes</td>
<td>≤ 0.06</td>
</tr>
<tr>
<td>From January 2008 to present</td>
<td>≤ 0.06</td>
</tr>
<tr>
<td>Meningitis, via intravenous administration</td>
<td>≤ 0.06</td>
</tr>
<tr>
<td>Non-meningitis syndrome,</td>
<td>≤ 0.06</td>
</tr>
<tr>
<td>via intravenous administration</td>
<td></td>
</tr>
<tr>
<td>Non-meningitis syndrome, Oral administration</td>
<td></td>
</tr>
<tr>
<td>Erythromycin</td>
<td>Any syndrome any route</td>
</tr>
</tbody>
</table>

In 2000 the PCV7 pneumococcal vaccine was added to the childhood vaccination schedule. In order to examine vaccination effects I analysed ABCS data collected before the year 2000 in the first instance. In further analysis I then repeated the analysis on isolates collected after the year 2000.

### 4.2.2 Duration of carriage data

I analysed data on the serotype specific duration of carriage which had been collected from the Kilifi Epidemiologic and Demographic Surveillance Study (Epi-DSS). Carried out in Kilifi, Kenya the Kilifi Epi-DSS is a longitudinal recording of residents in a well-defined geographic area around Kilifi District Hospital and is updated through continuous census for vital events every four months. Children younger than five years of age were randomly recruited from the Epi-DSS and a nasopharyngeal sample was collected with a swab at the beginning of the study. Children whose swabs were positive for pneumococcal carriage were re-sampled on approximately the 1st, 2nd, 4th, 8th, 16th and 32nd days from the first sample and monthly thereafter. Sampling stopped when they observed two consecutive swabs which were negative for the original serotype or negative for any pneumococcal growth (whichever came sooner) (208). The estimates of duration of carriage from this study are in Figure 4.1.
4.3 Description of statistical models and statistical techniques

From the mathematical model presented in Chapter 3, I observed a relationship between the duration of carriage and prevalence of resistance that can be approximately described by a sigmoid curve. A curve that can follow this relationship is described by equation 4-1. In equation 4-1 the prevalence of antibiotic resistance $f$ is determined by the mean duration of carriage $D$, dependent on the threshold parameter $D_{50}$ and parameters which govern the shape of the curve $\alpha$ and $\gamma$. The null hypothesis, where $D_{50} = 0$ and the prevalence of antibiotic resistance is independent of the duration of carriage is described by equation 4-2, where $\alpha$ is the mean prevalence of antibiotic resistance. To allow comparison with the parameters estimated from the mathematical models presented in Chapter 3, the parameter $D_2$ is calculated from the estimated values of $\alpha, \gamma$ and $D_{50}$, this is described by 4-3.

4-1 Statistical model

$$f(D) = \frac{\alpha D^\gamma}{D^{\gamma} + D_{50}^{\gamma}}$$

4-2 Null hypothesis

$$f = \alpha$$ or equivalently $$D_{50} = 0$$
4-3 $D_5$: Duration of carriage at which prevalence of resistance is 5%

$$D_5 = \left( \frac{0.05D_5\alpha}{\alpha - 0.05} \right)^{1/\alpha}$$

The null hypothesis is nested within the sigmoid model, thus both models are fitted to data using the maximum likelihood method.

Due to sampling effects, a small sample of a serotype is less likely to have resistant isolates present even if these resistant isolates are present in the population the sample comes from. Thus the likelihood calculations (shown in 4-4) include the frequency of each serotype as well as the prevalence of resistance from data and the model prediction. Using a ‘generalised reduced gradient algorithm’ optimisation algorithm, the parameters are varied until values are found which maximise the likelihood of the observed data. All analysis was carried out in Excel.

4-4 Likelihood equations

$$l_{\text{null}}(\alpha) = \sum_i N_i \ln(f) + (F_i - N_i) \ln(1 - f) \quad \text{(with 1 d. f.)}$$

$$l_{\text{sigmoid}}(\alpha, \gamma, D_{50}) = \sum_i N_i \ln(f^-(D_i)) + (F_i - N_i) \ln(1 - f^-(D_i)) \quad \text{(with 3 d. f.)}$$

$$l_{\text{saturation}}(N_i) = \sum_i N_i \ln\left(\frac{N_i}{F_i} \right) + (F_i - N_i) \ln\left(\frac{F_i - N_i}{F_i} \right) \quad \text{(with i d. f.)}$$

Where:

$N_i$, number of isolates of serotype $i$ penicillin resistant

$F_i$, frequency of serotype $i$

$D_i$, duration of carriage of serotype $i$

$f$, prevalence of resistance (null hypothesis)

$f^-(D_i)$, prevalence of resistance (sigmoid model) corresponding to duration of carriage $D_i$

Using maximum likelihood estimation, I estimated the parameters which determine the functional form of each model.

I then tested whether the second model described by 4-1 is a better fit to the data than the null model, 4-2. Using the likelihood ratio test (shown in 4-5), which is based on the value of the log likelihood ratio at the null value of the parameter, I tested the strength of the evidence against the null model. If the second model is a statistically significantly better fit than the null model, the corresponding $P$ value derived from the $\chi^2$ distribution with 2 degrees of freedom is $P < 0.05$. 
To assess the goodness of fit of the sigmoid model, the likelihood for the saturated model (which assigns for a parameter for each serotype) was calculated and compared to the sigmoid model using the $LRS$ as above. This result was statistically significant in all cases, indicating that the sigmoid model does not fully account for the variance seen in the dataset. Such unexplained variation is known as extra-Binomial overdispersion. Standard errors and therefore CIs from the sigmoid and null model were adjusted for overdispersion by using an overdispersion adjustment factor. This is the square root of the model deviance divided by the degrees of freedom.

A possible confounder of any relationship is the age of the patient, as all serotypes are not equally likely to cause invasive disease in each age group. To examine this confounder I fitted the model to samples from the children under two years of age only and then from samples of those aged over two.

Another important influence of serotype distribution is the implementation of the 7 valent pneumococcal carriage vaccine (PCV7) which provides immunity against colonization of certain serotypes: 4, 9V, 14 19F, 23F, 18C, 6B. This vaccine was included in the United States childhood vaccination schedule in 2000 and caused significant changes in serotypes circulating in carriage and causing invasive disease \cite{48,146,47}. To account for this I fitted the data against samples collected before 2000, and those collected after 2000.

### 4.4 Results

#### 4.4.1 Preliminary data analysis

Table 4.2 shows the frequency of each serotype in the ABCS dataset before 2000 and the estimated duration of carriage of each serotype estimated from the Kenyan nasopharyngeal carriage study. As this table shows there is a wide variation in the frequency that each serotype is isolated from invasive disease. Serotype 14 is the most common serotype isolated from both age groups, while some serotypes present at low frequencies in those over two (serotypes 35C and 2 for example) are not isolated from invasive disease in those under two.

Frequency distributions were calculated for penicillin and erythromycin susceptibility from the ABCS dataset. These were then stratified for age; under two years and over two years. Figure 4.2 shows the distribution of susceptibility to penicillin among each serotype of pneumococci isolated from
patients younger than two years of age (Figure 4.2 (A)) and older than two years of age (Figure 4.2 (B)). Figure 4.3 (A) and (B) presents the prevalence of erythromycin resistance in invasive disease occurring before 2000 in patients younger than two years of age Figure 4.3 (A) and those older than two years of age Figure 4.3 (B).
Table 4.2 Frequency and carriage duration of serotypes in patients under and over two years of age

<table>
<thead>
<tr>
<th>Serotype</th>
<th>Frequency</th>
<th>Mean Carriage duration (days)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt; 2 years</td>
<td>&gt; 2 years</td>
<td></td>
</tr>
<tr>
<td>19F</td>
<td>315</td>
<td>432</td>
<td>52</td>
</tr>
<tr>
<td>6A</td>
<td>149</td>
<td>502</td>
<td>46</td>
</tr>
<tr>
<td>6B</td>
<td>347</td>
<td>700</td>
<td>46</td>
</tr>
<tr>
<td>23B</td>
<td>7</td>
<td>17</td>
<td>33</td>
</tr>
<tr>
<td>11A</td>
<td>4</td>
<td>223</td>
<td>32</td>
</tr>
<tr>
<td>23F</td>
<td>224</td>
<td>865</td>
<td>31</td>
</tr>
<tr>
<td>35B</td>
<td>7</td>
<td>71</td>
<td>29</td>
</tr>
<tr>
<td>15A</td>
<td>3</td>
<td>23</td>
<td>28</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>10A</td>
<td>9</td>
<td>66</td>
<td>27</td>
</tr>
<tr>
<td>14</td>
<td>1027</td>
<td>1539</td>
<td>24</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>45</td>
<td>23</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>19A</td>
<td>89</td>
<td>335</td>
<td>19</td>
</tr>
<tr>
<td>35A</td>
<td>0</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>16F</td>
<td>5</td>
<td>66</td>
<td>18</td>
</tr>
<tr>
<td>18C</td>
<td>231</td>
<td>387</td>
<td>18</td>
</tr>
<tr>
<td>15C</td>
<td>11</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>15B</td>
<td>10</td>
<td>71</td>
<td>18</td>
</tr>
<tr>
<td>07C</td>
<td>0</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>004</td>
<td>200</td>
<td>1257</td>
<td>17</td>
</tr>
<tr>
<td>23A</td>
<td>4</td>
<td>58</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>552</td>
<td>11</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>65</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>302</td>
<td>9</td>
</tr>
</tbody>
</table>
Figure 4.2 Prevalence of penicillin resistance pre 2000 in invasive disease isolates from individuals (A) under two years of age and (B) over two years of age.

The serotypes are ranked from longest to shortest carriage duration from top to bottom. Green sections are the percentage sensitive (MIC break point of $\leq 0.06$ µg/ml), orange intermediate (MIC break point of 0.12-1 µg/ml), and resistant red. Resistance is defined as pneumococci with an MIC break point of $\geq 2$ µg/ml.
Figure 4.3 Prevalence of erythromycin resistance pre 2000 in invasive disease isolates from individuals (A) under two years of age and (B) over two years of age. The serotypes are ranked from longest to shortest carriage duration from top to bottom. Green sections are the percentage sensitive and resistant red, resistance is defined as pneumococci with an MIC to erythromycin of ≥1 µg/ml.
Penicillin resistance and erythromycin resistance are unevenly distributed among the serotypes. In pneumococci isolated from patients aged less than two years, serotypes 1, 15B and 15C are all susceptible to penicillin. Although more than half of pneumococci of serotype 14 isolated from children less than two years of age were resistant to erythromycin, no erythromycin resistance is reported in serotypes 1, 23A, 15A, 15B, 15C, 10A, 21, or 35B.

### 4.4.2 Statistical analysis

Using maximum likelihood estimation I analysed whether the null hypothesis (that the prevalence of resistance to antibiotics is randomly distributed among serotype) or the sigmoid model (a sigmoid curve) best describe the distribution of antibiotic resistance among serotypes. In the first instance I analysed pneumococcal isolates collected before the introduction of vaccination in 2000. Figure 4.4 (A) shows the prevalence of penicillin resistance in patients less than two years of age collected before 2000 plotted against the duration of carriage. The frequency of each serotype is plotted with its 95% CI, in the same figure are the null model (plotted as a dashed line) and the sigmoid model (plotted as an unbroken line) fitted to the data. The functional form of the relationship in the model has been determined through maximum likelihood estimation of the parameters as described in the previous section. The sigmoid model fitted a sigmoid functional form when fitted to data using maximum likelihood methods. Parameter estimates are presented in Table 4.3. From the null model, the estimated value \( \alpha \), of the mean prevalence of penicillin resistance is 0.25 (95% CI 0.18-0.23).

In the sigmoid model, \( \alpha \) is estimated to be 0.32 (95% CI 0.27-0.38), the parameter \( D_{50} \) 20.44 (95% CI 19.06-21.91) days. The sigmoid model (sigmoid curve) was found to fit significantly better to data than the null hypothesis and \( P < 0.001 \).

Although the sigmoid model fits significantly better than the null model, it does not fully account for variance in the data as the saturated model fits significantly better to the data than the sigmoid model. Therefore, all reported CIs are adjusted for overdispersion.

To examine whether the hypothesis held for a separate class of antibiotics, I repeated the analysis examining the distribution of erythromycin resistance among serotypes. Presented in Figure 4.5 is the prevalence of erythromycin resistance in pneumococci isolated from invasive disease in individuals under (A) and over two years of age (B), plotted against the duration of carriage. In both age groups the sigmoid model (a sigmoid curve) fits significantly better to data than the null hypothesis. Parameter estimates are presented in Table 4.4, in both age groups the prevalence of erythromycin resistance was higher than penicillin. In the analysis from the sample of isolates from
those under two, the value of the mean prevalence of erythromycin resistance ($\alpha$) estimated by fitting the null model was 0.32 (95% CI 0.23-0.42) and by fitting the sigmoid model was estimated as 0.42 (95% CI 0.34-0.51). The estimate of $D_{50}$ was slightly though not significantly lower than those values estimated for penicillin resistance.

The values of $D_5$ were calculated by using estimated parameters in the equation presented in 4-3. The value of $D_5$ for penicillin resistance was calculated as 18.78 days, and for erythromycin resistance, 16.39 days.
Figure 4.4 Prevalence of penicillin resistance. Mean proportion of serotypes penicillin resistant with 95% confidence bounds, plotted against the duration of carriage. A) Serotypes from surveillance of invasive disease in individuals under 2 years of age collected before 2000, B) Serotypes from surveillance of invasive disease in individuals over 2 years of age collected before 2000. The solid line plots the curve from sigmoid fitted to the prevalence of penicillin resistance (MIC ≥ 2), the dashed line, the null hypothesis.
Figure 4.5 Prevalence of erythromycin resistance.

Proportion of serotypes erythromycin resistant with 95% confidence bounds, plotted against the duration of carriage. A) Serotypes from surveillance of invasive disease in individuals under 2 years of age collected before 2000, B) Serotypes from surveillance of invasive disease in individuals over 2 years of age collected before 2000. The solid line plots the curve from sigmoid fitted to the prevalence of penicillin resistance (MIC ≥ 2), the dashed line, the null hypothesis.
Table 4.3 Maximum likelihood analysis: Association between prevalence of penicillin resistance (MIC ≥2) and serotype duration of carriage

<table>
<thead>
<tr>
<th></th>
<th>Pre 2000</th>
<th></th>
<th>Post 2000</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients under 2 years of age</td>
<td>Patients over 2 years of age</td>
<td>Patients under 2 years of age</td>
<td>Patients over 2 years of age</td>
</tr>
<tr>
<td></td>
<td>Null</td>
<td>Sigmoid</td>
<td>Null</td>
<td>Sigmoid</td>
</tr>
<tr>
<td>( \alpha ) (95% CI)</td>
<td>0.25 (0.18-0.33)</td>
<td>0.32 (0.27-0.38)</td>
<td>0.14 (09-0.21)</td>
<td>0.23 (0.18-0.30)</td>
</tr>
<tr>
<td>( D_{50} ) (95% CI)</td>
<td>20.44 (19.06-21.91)</td>
<td>20.61 (18.66-22.74)</td>
<td>18.84 (17.09-19.90)</td>
<td>18.84 (17.95-19.69)</td>
</tr>
<tr>
<td>( \gamma ) (95% CI)</td>
<td>19.98 (11.79-36.94)</td>
<td>15.53 (7.92-30.55)</td>
<td>26.85 (7.80-61.65)</td>
<td>55.72 (16.93-179.67)</td>
</tr>
<tr>
<td>( D_{5} )</td>
<td>18.78</td>
<td>18.94</td>
<td>17.58</td>
<td>18.38</td>
</tr>
<tr>
<td>( l_{\text{null}} )</td>
<td>-1506.14</td>
<td>-3101.45</td>
<td>-1179.56</td>
<td>-5090.84</td>
</tr>
<tr>
<td>( l_{\text{sigmoid}} )</td>
<td>-1369.94</td>
<td>-2648.93</td>
<td>-1064.16</td>
<td>4047.77</td>
</tr>
<tr>
<td>( l_{\text{saturated}} )</td>
<td>-1311.94</td>
<td>-2430.89</td>
<td>-918.42</td>
<td>-3402.52</td>
</tr>
<tr>
<td>( P \text{ value} \ (l_{\text{null}} \text{ and } l_{\text{sigmoid}} \text{ comparison}) )</td>
<td>( P &lt; 0.001 )</td>
<td>( P &lt; 0.001 )</td>
<td>( P &lt; 0.001 )</td>
<td>( P &lt; 0.001 )</td>
</tr>
<tr>
<td>( P \text{ value} \ (l_{\text{sigmoid}} \text{ and } l_{\text{saturated}} \text{ comparison}) )</td>
<td>( P &lt; 0.001 )</td>
<td>( P &lt; 0.001 )</td>
<td>( P &lt; 0.001 )</td>
<td>( P &lt; 0.001 )</td>
</tr>
</tbody>
</table>
Table 4.4 Maximum likelihood analysis: Association between prevalence of erythromycin resistance (MIC ≥1) and serotype duration of carriage

<table>
<thead>
<tr>
<th></th>
<th>Pre 2000</th>
<th>Post 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients under 2 years of age</td>
<td>Patients over 2 years of age</td>
</tr>
<tr>
<td><strong>Null</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$ (95% CI)</td>
<td>0.32 (0.23-0.42)</td>
<td>0.42 (0.34-0.51)</td>
</tr>
<tr>
<td>$D_{50}$ (95% CI)</td>
<td>19.52 (18.29-21.22)</td>
<td>19.65 (15.99-24.88)</td>
</tr>
<tr>
<td>$\gamma$ (95% CI)</td>
<td>17.70 (9.57-33.95)</td>
<td>8.87 (3.01-18.43)</td>
</tr>
<tr>
<td>$D_{5}$</td>
<td>16.39</td>
<td>17.55</td>
</tr>
<tr>
<td>$l_{null}$</td>
<td>-1364.97</td>
<td>-1364.97</td>
</tr>
<tr>
<td>$l_{sigmoid}$</td>
<td>-2460.309757</td>
<td>-1223.215859</td>
</tr>
<tr>
<td>$l_{saturated}$</td>
<td>-1147.72</td>
<td>-2225.52</td>
</tr>
<tr>
<td><strong>P value ($l_{null}$ and $l_{sigmoid}$ comparison)</strong>)</td>
<td>$P &lt; 0.001$</td>
<td>$P &lt; 0.001$</td>
</tr>
<tr>
<td><strong>P value ($l_{sigmoid}$ and $l_{saturated}$ comparison)</strong></td>
<td>$P &lt; 0.001$</td>
<td>$P &lt; 0.001$</td>
</tr>
</tbody>
</table>
4.5 Discussion and analysis limitations

I have shown that the distribution of penicillin and erythromycin resistant pneumococci among serotypes is associated with the difference in serotype specific carriage duration. The relationship between duration of carriage prevalence of resistance was described by a sigmoid curve with a threshold fitness advantage (duration of carriage) below which antibiotic resistance was not maintained in the population. This suggests that the assumption that genetic mechanism(s) of resistance are independent of serotype (59,212) is correct and the cost of resistance is constant in each serotype.

However analysis of antibiotic prevalence could be confounded by the impact of age or other effects. Exploration of the datasets revealed that serotype frequency varied among different age groups from the invasive disease isolates. The invasive disease dataset also showed serotype variation for different years due to the impact of a pneumococcal conjugate vaccination campaign.

i. Age

Certain serotypes are associated with infections in children. This age group is thought to be a higher consumer of penicillin and other antibiotics than older individuals (205). To examine whether age is a confounder I stratified each dataset into isolates from patients under and over two years of age.

The sample of pneumococci isolated from invasive disease in individuals over two years and analysed to examine the relationship between the prevalence of resistance and duration of carriage are presented in Figure 4.4 (B). Parameter estimates for the prevalence of penicillin resistance are presented in Table 4.3. In the sample of individuals older than two years, the value of the mean prevalence of penicillin resistance ($\alpha$) estimated by fitting the null model was 0.14 (95% CI 0.09-0.21). In the sigmoid model, ($\alpha$) 0.24 (95% CI 0.18-0.30), the parameter $D_{50}$ was estimated at 20.61 (95% CI 18.66-22.74) days. In this age group the sigmoid model fits significantly better to the data then the null model, $P << 0.001$.

Figure 4.5 (B) presents the sample of pneumococci isolated from invasive disease in individuals over two years of age and analysed to examine the relationship between the prevalence of erythromycin resistance and duration of carriage. Table 4.4 presents the result for erythromycin analysis. The value of the mean prevalence of erythromycin resistance ($\alpha$) estimated by fitting the null model was estimated at 0.23 (95% CI 0.22-0.24). In the sigmoid model, ($\alpha$) is estimated as 0.29 (95% CI 0.19-0.43) the parameter $D_{50}$ was estimated at 19.65 (95% CI 15.99-24.88) days. In this as in the younger age group the sigmoid model fits significantly better to the data then the null model, $P < 0.001$. 
As the analysis of both classes of antibiotics shows, in older patients the relationship between the duration of carriage and prevalence of resistance is described by a sigmoid curve. The prevalence of resistance is lower in this age group and the values of $D_{5}$ and $D_{50}$ are slightly though not significantly greater.

**ii. Pneumococcal conjugate vaccination campaign**

In 2000 the pneumococcal conjugate vaccine was introduced in the United States. This vaccine prevented carriage and invasive disease caused by serotypes: 4, 9V, 14 19F, 23F, 18C, 6B. This resulted in serotype replacement, where serotypes included in the vaccine declined in the population and are replaced by non-vaccine serotypes (48,47). Presented in Table 4.3 and Table 4.4 are the results of maximum likelihood analysis of isolates collected after the introduction of PCV7 vaccination in 2000. In the case of both penicillin and erythromycin in each age group the sigmoid fit significantly better to data than the null, and $P << 0.001$. Although there are some differences in the parameter estimates calculated from the samples pre and post vaccination, the CIs of these estimates are overlapping.

**iii. Dataset suitability**

In this analysis I used two datasets for the majority of analysis, the prevalence of antibiotic resistance per serotype was calculated from a dataset compiled from surveillance of invasive disease in the United States and the duration of carriage of serotypes was taken from estimates of duration of carriage of isolates colonising healthy children in Kenya.

The first and most obvious disadvantage of this approach is that these two datasets come from two very different settings. It is reasonable to expect that the prevalence of carriage and serotypes in circulation will differ between a high income country such as the United States and a low income country such as Kenya.

A comparable study of carriage duration from a high income country comes from the UK (31). The major advantage of the Kenyan dataset is its size and thorough sampling regime. Due to this the dataset has a larger range of serotypes (with clearly defined confidence intervals) than the UK dataset. The smallest sampling interval in the UK dataset was two weeks, while in Kenya in the first week, samples were taken at daily intervals. Therefore in the UK the shortest duration of carriage estimated was four weeks (for serotype 9A) and there was the potential for researchers to confuse multiple incidences of short carriage with one incidence of long carriage (31). For these reasons the Kenyan dataset remains the best dataset available for this analysis.
The second problem that arises in comparing the two datasets is that one dataset comes from invasive disease and the other from carriage. As is shown in Figure 4.6 (A) the prevalence of serotypes in carriage differs from that in invasive disease. While the ABCS dataset would be unsuitable to study serotype frequency, as serotypes which are common in invasive disease may be rare in carriage, there is no evidence to suggest that the prevalence of antibiotic resistance in invasive disease is not similar to that in the wider pneumococcal population.

Shown in Figure 4.6(B) the comparison of resistance prevalence in both datasets shows that the prevalence of resistance differs in the two datasets. However all isolates of serotypes 11A, 15B, 18C, 23A are not resistant to penicillin in both datasets. This indicates that although the prevalence of penicillin resistance may be sample dependent, the presence or absence of resistance in certain serotypes forms a consistent pattern.
Figure 4.6 Comparison of invasive disease and carriage datasets. A) Proportion of serotype from carriage and invasive disease. Proportion of each serotype isolate from a study of healthy children collected in 2001 (white column) (48). Black bars show the proportion of that serotype in isolates from invasive disease in children under two, collected by the ABCS programme US, 1996-2000 (216). B) Prevalence of penicillin nonsusceptibility in the carriage dataset (white column) and the invasive disease dataset (black column).
4.6 Discussion and conclusion

In this section I will examine some key insights from statistical analysis of invasive disease data, and interpret these with reference to predictions from the mathematical models presented in Chapter 3 and, where appropriate, previous studies in published literature.

i. **The prevalence of penicillin resistance varies among age groups**

From analysis of invasive disease data, the average prevalence of resistance was greater in samples from children younger than two years of age than from individuals over two years of age. This observation was true in the case of penicillin and erythromycin resistance. The most obvious explanation for this is increased antibiotic consumption in younger age groups. Surveys of antibiotic use in the community have shown that older age groups consume fewer antibiotics than children under two years of age (11). In both mathematical models presented in Chapter 3, increased antibiotic consumption resulted in increased prevalence of antibiotic resistance.

This finding is consistent with previously published mathematical models which examined the impact of treatment on resistance (82,80). The results are also consistent with studies examining the association between antibiotic consumption and the prevalence of resistance in pneumococcal populations. Goossens et al. (74) compared the outpatient consumption of antibiotics in 26 European countries with the prevalence of resistance in pneumococcal invasive disease. The authors used Spearman’s correlation coefficient and found a significant relationship between penicillin use and prevalence of penicillin resistance in IPD (74).

The results from the mathematical model presented in Chapter 3 and data analysis presented in Chapter 4 strongly suggest that differences in antibiotic consumption between those patients age under and over two years are responsible for difference in the prevalence of resistance between those two age groups.

ii. **Estimates of $D_5$ are consistent in different age groups**

The results of the mathematical model presented in Chapter 3 predict that different rates of antibiotic consumption will change the estimate of $D_5$. Decreasing the rate of antibiotic consumption from a 1/3 to a 1/6 per person per month increases the value of $D_5$ by over 10 days. In analysing the data in Chapter 4, I showed that although the prevalence of resistance decreased with age, the estimates of $D_5$ did not decrease with age to the same degree. This would suggest that although a decrease in the rate of antibiotic consumption is associated with a decrease in the average
prevalence of resistance within a sub-population, the impact on the distribution of antibiotic resistance among serotypes is not as pronounced. Therefore it would be therefore interesting to examine populations with universally lower and higher antibiotic consumption rates than those studied to examine whether estimates of $D_5$ from these populations varied as predicted by the mathematical model.

iii. **Estimate of $D_5$ can be used to estimate the cost of penicillin resistance**

This analysis initially sought to explain the heterogeneous distribution of antibiotic resistance among serotypes. From analysis of the mathematical model it became clear that there was a consistent association between antibiotic consumption $D_5$ and fitness cost of resistance. Thus, if the value of $D_5$ can be calculated from a sample of data this can be used to estimate the cost of resistance.

Previous research into the fitness cost of antibiotic resistance in the pneumococcus has focused on estimating the impact of resistance on the growth rate in culture (211). This research has showed that antibiotic-resistant pneumococcal strains grow slower than sensitive strains in the absence of antibiotics. A method that allowed us to estimate the fitness cost of resistance on transmission in natural rather than laboratory populations of bacteria would greatly enhance our study of pneumococcal antimicrobial resistance.

From the statistical analysis of IPD data I estimated a value of $D_5$. In analysis of penicillin resistance in children less than two years old $D_5$ was estimated to be 18.78 days. In both mathematical models presented in Chapter 3 the value of $D_5$ was estimated to range from nine to 35 days and the estimate was dependent on the cost of resistance and the rate of antibiotic consumption. Thus, if the rate of antibiotic consumption in a study sample was known, the suitability of the $D_5$ estimate from data could be assessed and the cost of resistance could be estimated.

A suitable estimate of the rate of penicillin consumption in children less than two years of age in the United States is not available. However, one study has estimated the rate of penicillin consumption in children under two attending a family practice in Massachusetts (205) and from the results of this study I calculated the antibiotic consumption rate of 1/3 per person per month. Although a local sample may not be completely representative of national trends, the rate of 1/3 per person per month is a plausible rate of consumption. For this rate of antibiotic consumption values of $D_5$ estimated from the mathematical model range from 9 to 38 days. The estimate of $D_5$ from invasive disease is within this range of parameters ($D_5 = 18.78$), which is consistent with a penicillin resistant strain being no less than 80% as fit as a penicillin sensitive strain.
4.7 Conclusion

In conclusion, from analysis of a large sample of pneumococcal isolates from invasive disease, I have shown that the relationship between the duration of carriage of a serotype and the prevalence of resistance (to penicillin and erythromycin) can be described by a sigmoid curve. This relationship is consistent in different antibiotic classes, and is not disrupted by the introduction of pneumococcal conjugate vaccination in the year 2000. Although there are differences in the prevalence of resistance between different age groups, (those under and over two years of age), the relationship between the serotype-specific duration of carriage and prevalence of resistance is consistently best described by a sigmoid curve. I also estimated threshold parameters which predict the duration of carriage at which prevalence of resistance is 5% ($D_5$), estimates of $D_5$ are relatively consistent between age groups, ranging between 18 to 20 days.
5 Selection and Recombination in Bacterial Populations: literature review

5.1 Introduction

In the previous chapters I focused on when an adaptation like antibiotic resistance is beneficial to a bacterium like *S. pneumoniae*. This chapter and the next are broader in scope and examine the impact of selection on population structure of a bacterial population and the role of homologous recombination.

There is a large body of experimental and theoretical work describing the process of selection which focuses on either sexual or asexual organisms. Classical population genetics has provided an excellent theoretical framework with which to understand evolution in populations of sexually reproducing diploids or asexual haploids (220). Unfortunately bacteria do not fit neatly into either of these two ideal types being neither truly sexual nor asexual organisms. Unlike asexual organisms, bacteria can transfer and reassort genetic material through homologous recombination (221). In sexual organisms recombination is an obligate part of their lifecycle but bacteria reproduce by binary fission, not having to recombine every generation and so cannot be considered truly sexual organisms (222).

As I introduced in the first chapter of this thesis, homologous recombination plays an important role in pneumococcal population genetics. Recombination occurs in three basic steps. Firstly, free DNA from another bacterium associates with the pneumococcal cell, secondly the free DNA enters the host cell cytoplasm and thirdly becomes incorporated into the genome (223). Recombination as described above results in replacement of a segment of the recipient chromosome with that from another bacterium of the same (or very similar) species. The size of the replaced DNA is thought to be small, but this is thought to vary according to the mechanism. The amounts replaced are not well established, although normally thought to be a few hundred base pairs of DNA to 30kb via transformation (14).

In its most common form recombination in eukaryotic populations is a symmetric process, where genetic material is exchanged between two genomes. Recombination occurs during meiotic reproduction in a diploid population as part of meiosis a duplicated paternal chromosome is paired with its duplicated maternal homologue. As the nucleotide sequences of the maternal and paternal chromosomes are similar they can undergo a recombination event which in this context is called a
chromosome crossover. This process produces two hybrids, the chromosome of paternal origin which has incorporated some maternal DNA and the chromosome of maternal origin which has incorporated some paternal DNA (224). Recombination can also occur in eukaryotic populations through gene conversion. This is an asymmetric process which occurs during meiosis in eukaryotic genomes. It occurs when a segment of DNA from one chromosome replaces the homologous section of DNA in the recipient chromosome. The donor chromosome remains unchanged while the recipient is now a hybrid chromosome (225,226).

In eukaryotic populations a recombination event occurs at least once per chromosomal replication (7) and is an obligate part of the lifecycle (226). Bacteria do not have to recombine as part of their reproductive process and so it is not an essential part of their life cycle. The relative contribution of recombination and mutation to the generation of genetic diversity in bacterial populations varies in different bacterial types. The variation in the ratio of recombination to mutation can result in the generation of distinct structures of diversity in bacterial populations. In populations with low levels of recombination the population is highly clonal, consisting of individually evolving lineages (227). An example of this is an organism such as *Mycobacterium tuberculosis*, which is highly clonal, and its population structure suggests that recombination plays little role in its diversification (228,229). In bacterial populations with high ratios of recombination to mutation, such as *Helicobacter pylori* (102) clones cannot emerge as their genomes diversify too rapidly. The majority of bacterial populations are thought to have rates of recombination that lie between these two extremes (227).

5.1.1 Aims

Incorporating recombination in population models is one of the challenges of bacterial population genetics. Another is to generate models which are flexible enough that they can simulate the range of recombination rates seen in bacterial populations. Therefore, there is a need to generate models of selection that incorporate the unique characteristics of bacterial reproduction in order to understand the structure and variation in bacterial populations. The aims of this chapter are to

- Review the literature from the study of selection in bacterial populations in the laboratory and in theoretical models,
- Review the literature from theoretical models which describe the impact of recombination during selection,
- Examine the differences between a model of a selective sweep which incorporates classical recombination and one incorporating unidirectional or asymmetric recombination as seen in bacterial populations.
5.2 Review of selection in bacterial populations

Laboratory populations of bacteria have been used along with other experimental systems such as fruit flies to test hypotheses about the selection process (230). Selection is thought to play an important role in shaping the genetic variation seen in natural bacterial populations. The role that selection is thought to play in bacterial populations has been greatly influenced by studies of laboratory cultures of bacteria undertaken since the 1950s (231,232). The term periodic selection is used in the literature to describe the process of the sequential occurrence and selection of adaptive bacterial types first observed in the continuous culture experiments of Atwood, Schneider and Ryan (233) and the chemostat experiments of Novick and Szilard (231,232). The term non specific selection was also used by Atwood to describe this process as the advantageous mutation occurred in an unknown gene that was not under study (233). However, periodic selection is the term most widely used to describe the phenomenon. The nomenclature is not ideal for two reasons. First, it suggests a distinct form of selection, when in fact it is used to describe a conventional process of asexual evolution which differs only in that the genes responsible for it are unknown. Second, the term implies a periodicity or regular recurrence of selective sweeps at predictable intervals, but this is not the case.

Important assumptions of the asexual model of selection are:

- As recombination is assumed to be absent, the advantageous mutation cannot be transferred onto another genetic background. This means that the whole genome is carried along with the advantageous mutant allele and the neutral markers on the genome used in experiments can represent the whole genome (234).

- Following from the first assumption, two beneficial mutations can only be incorporated onto the one genome if the second beneficial mutation occurs on the genome of the first (235). Therefore each adaptive clone is derived from the one preceding it in the population.

- A prediction that follows from the competitive exclusion principle is that when a population is growing under selection only one clone will be present at equilibrium (236). Therefore, other than during periods of adaptive shifts, when the parent population is being replaced by the mutant population, the population is assumed to be monomorphic.

Selective sweeps have been studied using laboratory models of bacterial evolution, mathematical models and a limited study of genomic data from natural populations of bacteria. Given that recombination is an important force in shaping the genetic structure of populations (227), the rest of
this chapter will review the study of simple selection in bacterial populations to ascertain whether the conventional asexual model is an adequate description of the selection process.

5.2.1 Review of previous studies of selective sweeps in laboratory bacterial populations

Laboratory experiments can be divided into those which used a batch culture technique (Figure 5.1) and those that use a chemostat. The chemostat differs from batch culture techniques in that it is an open system where medium is continually renewed and excess medium with waste products and dead cells are removed, shown in Figure 5.2. In the chemostat the population is maintained at a constant cell density, while in batch culture techniques populations of bacteria are allowed enter stationary phase, when the growth rate is null, before being transferred to fresh medium. Further detail and discussion of the laboratory techniques described in this section are provided in Appendix A.

Figure 5.1 Schematic diagram of batch culture technique.
First a sample of founder culture is added to a flask containing growth medium with suitable nutrients, populations of bacteria are allowed to grow until they enter stationary phase (when the growth rate is null) before being transferred to fresh medium. This is repeated for the desired number of generations.

i. Batch culture experiments

Atwood’s original experiments used serial transfers of *Escherichia coli* culture, histidine dependence was used as a neutral marker to detect the effect of the selective sweep. Histidine is an amino acid and histidine dependant bacteria cannot grow on media which does not contain a source of histidine. The mutation for histidine independence is neutral if the bacteria are grown in an environment where histidine is provided. During the experiment changes in the genetic structure of the population are monitored through observing the change in the proportion of cells that are histidine independent. A small proportion of the cells in the population are histidine independent, it
is assumed that it is very unlikely that a mutation would arise on a cell which is histidine independent (see the glossary in Appendix A for further details). This experiment found a selective sweep occurring after 200 generations when the frequency of histidine independent cells declined in the population. This is assumed to correspond with the occurrence of an advantageous mutant in the histidine dependant population of cells which then expanded through the population. The histidine dependent cells isolated during the experiment were, when competed against each other, identical in fitness. When competed against the histidine independent cells they were found to be competitively superior. When the sweep had come to completion all the histidine dependant cells were competitively superior to the original histidine independent cells. Histidine independent cells isolated after the sweep were selectively identical to the dependent cells, suggesting that these cells had been derived from the sweep population rather than remaining from the original population. Two more sweeps of advantageous mutants occurred in the population, each having a higher maximum growth rate than the preceding population (233).

These experiments were repeated using the same strains and media, though changing the culture conditions so that the experimental population alternated between growth and non-growth (237). The results of the experiment confirmed the findings of Atwood et al. (233) and showed that the adapted strains produced more acetate during growth. The authors then mapped genetic changes to the region containing genes responsible for the production of pyruvate dehydrogenase. They hypothesised that the glucose limited environment selected for mutations which allowed an enhanced ability to metabolise glucose efficiently.

The experiment was repeated again but using different strains of *E. coli* and different neutral markers, and found a selective sweep occurred after 400 generations (238). Lenski et al. carried out a second serial culture experiment to investigate the long term trajectory of mean fitness in a population undergoing periodic selection (239). This was a long term serial culture experiment with limiting glucose, which was an environment that the bacterial cells were ill adapted to, with 12 replicate cultures run for 2000 generations. In this experiment clones were not isolated at the end of each turnover but rather isolated every 100 generations. This method tracked changes in fitness over the 2000 generations. The relative fitness of the competing clones was estimated by performing a fitness assay (further described in the glossary in appendix A). In this case the clones were distinguished by their ability to grow on arabinose (a monosaccharide) and were distinguished by being grown on a specialist agar which resulted in the competitors forming different coloured colonies. Each clone was competed against the original strain, and it was found that the mean fitness of the isolated clones increased over time. The mean rate of the increase in fitness was calculated.
using a least squares linear regression of mean fitness against time for each of the 12 populations over 2000 generations and over the first 1000 generations. The mean rate of the increase in fitness declined with time (239). There was no evidence found of fitness variation within the populations. There was no investigation as to how the advantageous mutants increased the fitness of the cells or any analysis as to the genomic changes occurring in the population. This limits our understanding of the impact of selection in the population. It was hypothesised that the increased fitness was due to a shorter lag phase on being transferred onto fresh medium, reduced mortality during stationary phase and greater affinity for limiting glucose (239). The study provided evidence for the populations converging on a single adaptive peak, where fitness of the clones increase over time, so that each new mutant is more fit that the original clone. At each time point tested, all isolates tested were equally fit, suggesting they were a homogenous population of clones.

Early experiments performed using serial transfers of batch culture consistently show a characteristic pattern of multiple sweeps with transitive increasing relative fitness (238). Fitness is homogenous within populations outside of the periods of adaptive shift. The studies support the classical asexual model of selection but do not investigate the genetic structure of the population. More recent research by Rozen and Lenski found that two phenotypically distinct types of bacteria, evolved from a genetically homogenous starting culture and persisted in a stable equilibrium for 20,000 generations, providing clear evidence for the maintenance of polymorphisms in *E. coli* under glucose limitation (240). This contradicts the classical model which predicts that only one type of clone will be present at equilibrium in the population.

The two methods of characterising the population undergoing selection used in the previous studies - monitoring the neutral marker and using fitness assays to determine the relative fitness of the mutant to the ancestral strains - can provide good evidence that change is taking place in the population, but are limited in their ability to characterise that event (241). Monitoring the frequency of a neutral marker does not allow the researcher to determine the origin of each new advantageous mutant. The data from past experiments (233,237,238,240,241) cannot rule out a more complicated model than the classical one which would allow non sequential mutation. In the multiple adaptive mutants model there is a reservoir of adaptive mutations in the population at varying frequencies (242). There will still be adaptive shifts in these populations but not necessarily in a sequential order, so the fourth advantageous clone may not contain the mutation that characterised the third dominant clone.
ii. **Chemostat experiments**

In the first chemostat experiment demonstrating periodic selection performed by Novick and Szilard, the first selective sweep occurred after thirty generations (231). This was in a chemostat environment where tryptophan was the limiting nutrient. The experiment was run for 500 generations and six sweeps occurred during that time. In the experiment which used serial transfers of batch culture the first sweep did not occur until after 200 generations (233). The relative fitness coefficients of the two experiments also differed when estimated indirectly from the experimental data (rather than using fitness assays) (243). The selection coefficient for the advantageous mutant that emerged in the serial culture experiment of Atwood et al. was 0.37 and the selection coefficient for the first advantageous mutant occurring at thirty generations in the experiment of Novick and Szilard was 0.95 when estimated indirectly from the published data (231,233,243). Kubitschek postulated that the *E. coli* populations which were used in both experiments, though from different original cultures, were better adapted to the “feast and famine” environment of glucose-limited serial culture than the continuous growth environment of constant tryptophan limitation in the chemostat experiment of Novick et al. (243). A shorter time until the first sweep and increased fitness coefficients has been seen consistently in chemostat experiments, suggesting that the chemostat environment in which bacteria are growing continuously exerts much greater selective pressure than the batch culture environment. It was found that the advantageous mutants were more efficient at utilising substrate than the original population, without decreasing their maximal growth rate. In Dykuizen and Hartl’s experiment, in glucose-limited medium chemostat that was sampled every 100 hours, an increase in fitness was observed every 100 hours. In this experiment the selection was for isolates that utilised the chemostat nutrients more efficiently (244).
More recent work has identified stable polymorphisms which evolve repeatedly in originally homogenous populations growing on a single limiting resource. A genetically homogenous population of *E. coli* was shown by Rosenzweig et al. to evolve into a stable three clone polymorphic population. The study found evidence for the evolution of resource partitioning as a result of adaptations in the acetyl co-enzyme A synthase which allowed the co-existence of acetate scavenger bacteria and bacteria with a superior rate of glucose uptake (245). Further study has shown the repeated evolution of the acetate cross feeding polymorphism (246). Other work has confirmed the evolution of stable polymorphism in chemostat populations. Recent work has shown consistent evidence for the maintenance of polymorphisms in *E. coli* under glucose limitation (247-252). These studies have shown adaptive mutations in the *mgl, mlc* and *malT* loci (see the glossary in Appendix A for a brief explanation of gene cluster names) and that adaptive populations acquired multiple adaptive alleles rather than one single winner clone (40-45). In the experiments they used the traditional method of studying fluctuations in a neutral mutation which confers resistance to T5 phage infection, but they also sequenced the genes on which they believed the adaptive mutation to be occurring - *mgl, mlc* and *malT*. The results of the experiments showed that although shifts in the T5 resistant population could be interpreted by the classical model as a series of selective sweeps, when the loci of interest were sequenced it was found that genotypic diversity was maintained in
the population, thus demonstrating that the classical model was incorrect. In the study by Majaran et al., it was found that a polymorphism existed where changes in the outer membrane proteins were detectable so that the bacterial population could be subdivided into five distinct groups. Sequencing also showed multiple branched clusters in the population (252).

The studies of evolution in the laboratory described in the previous paragraph have revealed that the classical model of selective sweeps may not always hold, not even in the relatively simple environment of the chemostat or batch culture. They reveal much more genetic diversity than was previously seen in the original experiments in the 1950s. The more recent experiments have used genomic and biochemical techniques to characterise the population in the experiment while the earlier experiments used fitness assays and tracked the proportion of a neutral marker in the population. In experiments such as that by Dykuizen and Hartl (253), a detected change in fitness is equated with the replacement in the population of one bacterial type for another. This assumption leads to the conclusion that if a population is monomorphic in terms of fitness then there must only be one type of bacteria in that population. This is problematic as the types of fitness assays used in these experiments do not detect small changes in fitness and assume that if there is no detectable difference in fitness between two isolates then there is no genetic or phenotypic difference between those two isolates. In fact the two isolates could differ, but the difference in fitness between them is too small to be detected by the assay.

Two assumptions underpin the experiments which used a neutral marker to track changes in the population. Firstly they assume that one gene can represent the whole genome, and secondly that a relatively small assayed proportion of the population is representative of the genomic structure in the whole population. The studies described above showed that just tracking the neutral marker did not reveal the diversity that was found when sequencing was carried out (247-252).

All the experiments above assume that the bacterial populations are effectively asexual and that the neutral marker can be representative of the whole genome. No investigation took place considering whether the advantageous or the neutral marker could be separated from the genome on which it arose, or the role of recombination in the population. This is a valid assumption in these systems as there are no known mechanisms by which the chosen E. coli strains can exchange genes. In other bacteria this assumption is not valid (254).

5.2.2 Mathematical models of selective sweeps in bacterial populations

The way that bacteria in natural populations are distributed through space and time is likely to be different to that seen in the laboratory. Mathematical models provide an opportunity to model
selective sweeps in a variety of conditions that may provide insight into the role of periodic selection in natural populations of bacteria.

The earliest mathematical depictions of the periodic growth curves observed in a chemostat were presented by Koch (255). Who constructed a deterministic model and investigated the ranges of selection and mutation parameters which generate periodic growth curves (255). It considered two subpopulations which have different growth rates. A deterministic model found parameter values for the selection coefficient which allowed the best fit of the model to dynamics seen in the laboratory. Using the model Koch found that an increase in the abundant type is independent of the mutation from rare to abundant type, as the amount of rare type compared to abundant type in the population is always too small to make a substantial difference. The mutation rate and selective advantage of new mutants were parameters that made a substantial difference. This model describes the fluctuation in neutral alleles as seen in the chemostat but does not attempt detailed analysis of the occurrence of mutations, their selective advantage, or the effect of the sweep on the genetic structure of the population.

Deterministic models are not necessarily suitable for studying a process such as the recurrence of sweeps which may have a stochastic component such as the time before a sweep can occur. Berg constructed a model where the deterministic growth of a mutant variant is preceded by periods of stochastic delays in the appearance of novel advantageous mutants (256). In the model the early phase of appearance and growth of a new variant is modelled stochastically while the later phases when the new variant is at a sufficiently high number of cells is modelled deterministically. The model only considered the appearance of new variants that leave descendents. The time of appearance of a new variant was calculated from the mutation rate constant, the size of the population of the parental variant and the survival probability. This gives a distribution for appearance time. This was used to examine the impact of the stochastic delays on the fit of the simulated growth curves to experimental data. It was also used to calculate the hitchhiking probability, which is the probability that a neutral allele will be linked with and carried to fixation with an advantageous mutant. In this model hitchhiking by a weak or counter selected mutant takes place when the advantageous mutation occurs on the genome containing the weakly or counter selected mutant. The model assumed that bacterial growth is clonal and does not allow for the impact of recombination.

The effect of periodic selection on *E. coli* population structure was investigated by Levin in a mathematical model which examined the effect of periodic selection on effective population size. The effective population size of a population is the number of individuals in a theoretically ideal
population that have the same magnitude of random genetic drift as the actual population. The actual size of a population (the number of individual bacteria in that population) can increase during a selection event. Therefore a clonal population of bacteria could be large in number but have a low amount of diversity and consequently a low effective population size. Levin’s model examined the relationship between rare recombination, periodic selection and the effective population size (257). Levin investigated whether rates of recombination in *E. coli* populations were sufficiently high to override the bottleneck effects of periodic selection and afford these populations relatively large effective population sizes. He found that the effective population size increased exponentially as the recombination rate increased. The model used parameters estimated for natural populations of *E. coli* and found that the reduction in genetic diversity on the genome as a result of selection was substantial. With these parameters populations of *E. coli* would have very small effective population sizes. Levin simulated the model using a narrow range of parameters which had been mostly estimated from experimental rather than natural populations of bacteria.

The three models described in the preceding paragraphs have all addressed periodic selection in a chemostat or batch culture and attempted to model the observed growth curves (255,257,256). Majewski and Cohan’s model of bacterial evolution looks at the effects of global and local sweeps in a metapopulation structure (258). It examines the role of recombination within and between subpopulations and uses a coalescent model to simulate the evolution of a population into distinct sequence clusters called ecotypes, which constitute distinct ecological groups. Simulating the emergence of locally and globally beneficial mutations, it examines the potential of globally successful strains to sweep through the total population. When a globally beneficial mutant arises, it has the potential to reduce diversity in the whole population if the associated genome hitchhikes with the beneficial mutation.

Based on their model Majewski and Cohan found that with low recombination rates globally adapted mutants reduce neutral sequence divergence between ecological groups, but that this is dependent on the size of the fragment transferred at each recombination event. The larger the recombining fragment, the more effective the global sweep will be at reducing between population divergence. However, the model results provided evidence that global periodic sweeps would be able to homogenize subpopulations so that they are indistinguishable. When a global selection event occurs, the recombination rate between subpopulations needs to be above $10^{-7.8}$ for two subpopulations to homogenize, even when the recombining fragment size is greater than 10% of the genome. This threshold of recombination is higher than that currently estimated in most bacterial populations.
Central to the model is the assumption that periodic selection is an important evolutionary force in natural bacterial populations (259). As a result, the model fails to examine closely the effect of a selective sweep on a population for example the relationship between selective advantage, recombination rate and fixation time. This model of bacterial evolution has been suggested as one in which the purging effect of sweeps can be reconciled with the genetic diversity in bacterial populations. It has been suggested that bacteria live in distinct ecological niches and that locally advantageous mutations sweep to dominance within an ecotype or niche. Therefore, within a niche there is little diversity, but an abundance of niches allows the persistence of diversity over the wider bacterial metapopulation (258).

A recent model of bacterial evolution developed by Levin and Cornejo, was used to explore the contribution of homologous recombination to the rate of adaptive evolution in bacterial populations and the conditions under which homologous recombination will provide a bacterial population a selective advantage over non-recombining or more slowly recombining populations (260). Using a semi-stochastic model of bacterial evolution, results of their simulations indicated that under broad conditions recombination occurring at rates in the range anticipated for bacteria like *S. pneumoniae*, *E. coli*, *Haemophilus influenzae*, and *Bacillus subtilis* will accelerate the rate at which a population adapts to environmental conditions. The authors measured the adaptation to environmental conditions as an increase in mean fitness (260). Though this study provides a mathematical counterpart to laboratory studies on the impact of recombination on population fitness, it was not concerned with the impact of recombination on diversity and divergence in the population.

Like the previous model and simulations Johnsen et al. examine the impact of recombination on population fitness. The focus of the research was to examine why recombination is maintained in bacterial populations even though it is thought to carry a risk or fitness cost. The authors combined observations of growth in *Bacillus subtilis* populations with a mathematical model which simulated the population dynamics. From these observations they formulated the theory that recombination can be maintained in a population due to periods of non-growth when environmental factors (such as antibiotics) inhibit dividing cells. In *Bacillus subtilis* competent cells (cells that can take DNA from the surrounding environment into the cell to recombine with its genome) are non-growing. When attacked by an antibiotic these cells are not killed and so have a selective advantage over growing cells which are killed. Thus the recombining cells are maintained in the population (261).
5.2.3 Eukaryotic models of recombination and selection

The impact of recombination on the genetic diversity of a population undergoing a selective sweep has been studied comprehensively within eukaryotic population genetics. Theoretical models (262,263) and observations of the process in genomic samples (264) have led to a good understanding of the dynamics of selective sweeps in obligate sexually reproducing eukaryotes.

A selective sweep was shown to have an impact on patterns of linked genetic variation and the phenomenon known as the hitchhiking effect. This was demonstrated with a simple two locus haploid model (developed by Maynard Smith and Haigh) where symmetric recombination (the reciprocal exchange of genetic information during a single cross over event) occurred at a variable rate (265). This model demonstrated that when there are low rates of recombination between a neutral and selected allele a neutral allele may sweep to fixation with the linked beneficial allele. The genomic pattern predicted by this model has been observed in eukaryotic genomes where a selective sweep may result in previously rare alleles becoming very common. The model can also be used to show that recombination can allow neutral alleles to “escape” a sweep: they remain in the population even if they are not linked to the selected allele.

More recent work has concentrated on the effect of symmetric recombination on local linkage disequilibrium (LD). McVean et al. described LD between neutral alleles at two loci as arising because of correlations in the genealogical history of the two loci. Under the coalescent model, if the time to most recent common ancestor (MRCA) at locus x on the genome is informative about the MRCA locus y then the alleles at those loci should show significant LD (266). This study showed that a measure of LD was dependent on the correlations between coalescent time between two pairs of loci, which in turn depends on the rate of recombination between two loci (266). This result was used to examine the fate of two neutral loci situated near a locus that has undergone a selective sweep. The theory was validated using stochastic simulations and showed that the effect of a selective sweep on patterns of genomic diversity (measured by LD) was highly dependent on the position of the neutral allele simulated relative to the selected gene. In fact, depending on the relative position of the loci a selective sweep could increase, decrease or eliminate linkage equilibrium (263). Though the theory provided a comprehensive examination of the effect of symmetric recombination on a eukaryotic genome it failed to examine the impact of asymmetric recombination in the form of gene conversion which has been demonstrated to play a role in eukaryotic genomes. In the original model one of the results showed that two loci on the same side of the selected locus may show substantial LD whereas the expected LD on the opposite side of a selected locus is zero. When Wakefield et al. (267) extended the model to incorporate gene
conversion they showed that there may be substantial LD between two loci on opposite sides of a selective sweep (267). These models provide a robust explanation of genetic variability in eukaryotes although its predictions are not necessarily applicable to bacterial populations as they don’t explicitly model the effect of asymmetric recombination alone on the population.

5.3 Two-Locus Model of a selective sweep with asymmetric recombination

Although there has been substantial theoretical study of the role of symmetric recombination and the concurrence of asymmetric recombination, this has previously been carried out under the assumption that the recombination looked and behaved like that in a eukaryotic genome. Homologous recombination in bacteria differs from what is referred to as recombination in sexual eukaryotes in that it involves small regions of the genome, is not obligate with reproduction and rather than being a symmetric exchange of genetic material, it is a unidirectional transfer in which a small region of the chromosome of a recipient is replaced by homologous recombination with the corresponding region of a donor bacterium. In fact it is more similar to gene conversion as described in the model of Jones and Wakeley (267).

In this section I will present the comparison of two locus models of a selective sweep in a haploid population. The first (developed by Maynard Smith and Haigh) incorporates asymmetric recombination (265). The second which I developed, differs from the earlier model of Maynard Smith and Haigh by incorporating asymmetric recombination. I will examine whether the mechanistic differences between asymmetric and symmetric recombination impact on the results predicted from a simple two locus model. The advantage of using a simple model to do this is that firstly the model structures can be easily compared by comparing differential equations and secondly the impact of parameter variation can be easily examined.

The model in this chapter is developed as a ‘sample model’ to compare with previous work on this topic. I developed it to examine the impact of asymmetric compared to symmetric recombination and in the next chapter of this thesis I will use this model to examine the impact of asymmetric recombination on a selective sweep.

Like the model of Maynard Smith and Haigh I simulated a simple population where genomes are defined at two loci (265). The model is parameterised by a variable recombination rate, selection coefficient and population size. Under the infinitely many alleles model (IAM) mutation always generates new alleles. Alleles which become extinct as a result of a selective sweep cannot re-occur in the post-sweep population through mutation. Thus, pre-existing alleles can only remain in the population if they are rescued into the genotype sweeping to fixation through recombination. I
adapted the simple two locus deterministic model to incorporate an asymmetric recombination process as occurs in bacterial populations.

5.3.1 Methods

i. Model Structure

The parental population of the model consists of 1 pure mutant bacterium which carries an advantageous mutant allele (M) at the locus under selection and N-1 bacteria which represent the pure wild-type population, where N is the population size and remains constant. Under the two locus model the bacteria have one locus on which selection acts, with alleles M for the mutant genotype and W for the wild-type, and one neutral locus with two alleles X and Y. The fitness of W is 1 and the fitness of the M allele is 1 + s, where s is the selection coefficient. In the initial population, which is illustrated schematically in Figure 5.3 (Pre-sweep population), the neutral allele X is exclusively associated with the wild-type allele W and allele Y is exclusively associated with the advantageous mutant allele M. I refer to the MY bacteria as the pure mutant and the WX bacteria as the pure wild-type. Recombination allows four possible combinations of alleles, WX and MY which are present at the beginning of the simulation and WY and MX hybrids which are created only through recombination.

The frequency of each bacterial genotype at each generation of the model is parameterised by the selection coefficient $s$ and the population recombination rate $\rho$. The deterministic model with asymmetric recombination is described by the equation groups 5-1 to 5-3. Group 5-1 describes the proportion of each genotype at each generation, equation group 5-3 describes the definition of the per bacteria recombination rate $r$ used in the deterministic model defined in terms of the population recombination rate $\rho$.

In group 5-3 the proportion of each genotype in the next generation is determined. Under selection the frequency of genotypes that carry the M allele are multiplied by (1+s) and those that carry the W allele are multiplied by 1. During the recombination step a bacterium such as WX can be in the population already and not recombine, it can recombine with either a W or X allele and so remain unchanged in the population or recombine with an allele other than W or X resulting in a different bacterial genotype and the amount of WX in the population decreases. WX can increase in the population if a bacterium which contains either the W or X allele (i.e. WY or MX) recombines to create a WX genotype.

During a selective sweep, the descendents of the pure mutant MY increase in number until at the end of the simulation all bacteria have the M allele, as seen in Figure 5.3 (Post-sweep population). In
the two locus model there are two possible genotypes of bacteria in the Post-sweep population presented in Figure 5.3: MY which is the pure mutant genotype or MX which is a hybrid of the pure wild-type (WX) and the pure mutant genotype (MY). In this model the only way for a wild-type neutral allele (allele X in this model) to remain in the population after the sweep has occurred is through recombination with the pure mutant genotype, so that it becomes rescued through association with the M allele. The wild-type neutral allele X can recombine onto a genome containing the advantageous mutant allele M or vice versa. Both result in the X allele being retained in the population.

Figure 5.3 Two locus model of selective sweep.
A) Pre-sweep population. Initial population is composed of two genotypes. B) Recombination, recombination can occur at rate $\frac{D}{2N}$ between the loci in either direction. C) Selection each bacterium has two loci, locus one on which selection acts with alleles W or M and the second, the neutral locus with alleles X or Y. W has a fitness of 1 and M has a fitness of $(1+s)$ where s is the selection coefficient. The fitness of the hybrid type is dependent on the fitness of locus one. D) Post-sweep population, at the end of the sweep all genomes contain the M allele (red) at locus one.
\[ P_{w}(t) = \frac{N_{wx}(t) + N_{wy}(t)}{N} \]

\[ P_{m}(t) = \frac{N_{mx}(t) + N_{my}(t)}{N} \]

\[ P_{s}(t) = \frac{N_{sx}(t) + N_{sm}(t)}{N} \]

\[ P_{y}(t) = \frac{N_{wy}(t) + N_{my}(t)}{N} \]

5-2

\[ r = \frac{\rho}{2N} \]

5-3

\[ P(WX)(t + 1) = \frac{N_{wx}(t)[1-2r+P_{w}(t)+rP_{y}(t)]+N_{wy}(t)rP_{y}(t)+[1+s](N_{mx}(t)+N_{my}(t))}{(N_{wx}(t)+N_{wy}(t))(1+s)} \]

\[ P(WY)(t + 1) = \frac{N_{wy}(t)[1-2r+P_{w}(t)+rP_{y}(t)]+N_{wx}(t)rP_{y}(t)+[1+s](N_{mx}(t)+N_{my}(t))}{(N_{wx}(t)+N_{wy}(t))(1+s)} \]

\[ P(MY)(t + 1) = \frac{N_{my}(t)[1-2r+P_{m}(t)+rP_{y}(t)]+N_{wy}(t)rP_{y}(t)+[1+s](N_{mx}(t)+N_{my}(t))}{(N_{my}(t)+N_{wy}(t))(1+s)} \]

\[ P(MX)(t + 1) = \frac{N_{mx}(t)[1-2r+P_{m}(t)+rP_{x}(t)]+N_{my}(t)rP_{x}(t)+[1+s](N_{mx}(t)+N_{my}(t))}{(N_{mx}(t)+N_{my}(t))(1+s)} \]
ii. **Simulation**

Models were programmed using C++. The population size N was \(10^6\). Values for the population recombination rate \(\rho\) from 1 to 10000 were simulated. Values for the selection coefficient \(s\) were simulated from \(10^{-3}\) to 0.9 so that the relative fitness of the mutant allele was varied from 1.001 to 1.9.

5.3.2 **Results**

i. **Impact of recombination on pre-sweep allele prevalence**

Figure 5.4 shows a comparison between the results of the two locus model of a selective sweep with symmetric recombination (265) called Sym and the two locus model incorporating asymmetric recombination called Asym in Figure 5.4. I present the results of the two models simulated for selection coefficients 0.001, 0.01, 0.1 and 0.9. In these simulations the fitness of a bacterium carrying the advantageous allele M would be 1.001, 1.01 and so on and the fitness of a bacterium carrying the wild-type allele W at that locus would be 1. I also simulated the model for population recombination rates from 0.01 to 1000.

Plotted in Figure 5.4 is the proportion of genomes in the post-sweep population which carry an allele from the wild-type population, that is the proportion of the population that are hybrids carrying an M and an X allele. The MX genotype is not present in the pre-sweep population and can only be created through recombination. As shown in Figure 5.4, the proportion of alleles from the wild-type population that are rescued is low for a high selection coefficient and low recombination rate. When the population recombination rate is lower than 10 and the selection coefficient greater than 0.001, less than 10% of the final population is composed of bacteria that contain an allele from the wild-type population. As the selection coefficient gets smaller the number of alleles “rescued” through recombination from the wild-type population increases. This is because the smaller the selection coefficient the greater the amount of time that the WX bacteria remain in the population thus increasing the number of recombination events between an X allele and a MY bacteria or an M allele and a WX bacteria.
Figure 5.4 Comparison between Symmetric and Asymmetric models. This plots the relationship between the population recombination rate, selection coefficient and proportion of MX in the final population in the two locus model. The log of population recombination rate $\rho$ is plotted against proportion of MX in the final population when a sweep has occurred. The results are shown for fitness values of $1+s$, where $s$ is 0.001 (red lines), 0.01 (black lines), 0.1 (green lines) and 0.9 (blue lines). The results from the symmetric model are plotted as dashed lines, those for the asymmetric model as solid lines.

Figure 5.4 shows a comparison between the results of the two locus model of a selective sweep with symmetric recombination (48) and our two locus model incorporating asymmetric recombination. As shown in Figure 5.4, when the $\rho$ is lower than 10 and $s$ greater than 0.01, less than 10% of the final population is composed of bacteria with genotype MX. As $\rho$ increases the amount of MX in the final population increases, and as $s$ increases the amount of MX in the final population decreases. This pattern is simulated by both models.

However though both models follow the same trends in response to changes in the $\rho$ and $s$ there are numerical differences in the two results as demonstrated in Figure 5.4. At higher values of $s$ (0.01, 0.1 and 0.9) and with $\rho \geq 10$, the results from the symmetric model predict that the proportion of MX remaining in the population will be lower than if the model incorporates asymmetric recombination. This is also true for simulations when $s = 0.01$ and $\rho \leq 10$. 

121
The important structural difference between the earlier eukaryotic model and the asymmetric model used here is the recombination process. The differences in the mathematical model results show the importance of considering the biological differences in the recombination processes when studying bacteria. Rather than reciprocal genetic exchange, homologous recombination in bacteria involves the introduction of DNA from a donor into a recipient and replacement of a small region of the recipient with that from the donor—a process that has been described as localised sex. The effect of the difference between asymmetric and symmetric recombination can be most clearly illustrated by considering recombination between two bacteria that are identical at one locus. In the symmetric model the progeny of a recombination event between a MX and a MY bacteria will be MX and MY. Thus the amount of MX and MY will not change in the population with or without a recombination event. In an asymmetric recombination event, an allele from the donor bacteria replaces the allele at the appropriate locus in the recipient. When MY “donates” M and Y alleles, the amount of MX in the population will remain the same if a bacteria receives an M allele but will decrease if it receives a Y allele. While the difference in predictions will be small at the two locus level, these would increase in importance if many loci were considered. Thus even in a simple two locus haploid model bacterial recombination must be considered as a distinct event from that of eukaryotic recombination and modelled as such.

5.4 Conclusion

I conducted a review of the literature to examine findings from bacterial laboratory populations and mathematical models of bacterial and eukaryotic populations on the impact of selective on genetic structure.

The bacterial population studied in the laboratory were laboratory adapted *E. coli* (233,234,238). These bacteria are non-transformable, that is they do not take up DNA from the surrounding environment easily. Therefore the assumption that recombination would not occur in these populations is reasonable. These populations can be considered to be asexual and that any genetic change in these populations would arise from mutation not recombination. In early studies the mutation that conveyed the advantage to a mutant is uncharacterised and its prevalence is tracked in the population through the decline of a population which carry a gene that conveys a testable phenotype. These studies showed that a genotype which carries an advantageous mutant increased in the population and that the population was homogenous at the end of a sweep. These studies conform to the asexual model of selection. Later laboratory studies characterised the study population using genetic and biochemical tests (245,240,248-251,268,252,269). These studies revealed a more complex genetic structure than that predicted by the asexual model. In these
populations authors maintained that recombination could not occur and that the bacterial population had diversified in the chemostat environment as the ecotype model predicts would occur in an environment with multiple niches.

Laboratory populations of bacteria are very different to those seen in nature. A particular difference of interest is the lack of recombination in species of bacteria studied in the laboratory (220). Mathematical models provide the ability to examine the impact of particular parameters such as recombination or mutation on selection. Mathematical models of bacterial populations were initially formulated to simulate selection in a chemostat environment. They simulated an asexual population and the results of those models were consistent with the classical asexual predictions of selection (255,257,256). Like the classical asexual model they predicted that when an advantageous mutation arises on a genome descendents of that genome will sweep to fixation in the population. More advanced models of selection included recombination in bacterial populations. These models were concerned with examining the impact of selective sweeps on overall genetic diversity and did not examine directly whether recombination can retain neutral alleles from the wild-type population (257,258).

The predictions from eukaryotic models are more complex. Depending on parameters used, the simulations produced by the models predict that the effect of a sweep on measures of linkage disequilibrium in a population is very dependent on the position of the linked neutral allele under study. Broadly alleles near to the selected locus may hitchhike with the advantageous mutant to fixation but neutral alleles that are further away on the genome behave as though they are unlinked to the advantageous allele. Therefore diversity at some point on the genome may be conserved (263,267). Bacterial genomes are much smaller and differ structurally from eukaryotic genomes and as I outlined in the introduction there are significant structural differences between haploid symmetric recombination and bacterial asymmetric recombination (270). These structural differences mean that there are differences in the predictions of the two models. Although mathematical models incorporate a process like bacterial recombination into the eukaryotic model, these models still examine the effect selection in a eukaryotic population. Eukaryotic populations have both types of recombination occurring in tandem, whereas in bacterial populations only asymmetric recombination occurs at variable rates in different populations. There remains the need to use a mathematical model of selection in a bacterial population that can examine the effect of recombination on the selective sweep.

The preliminary predictions of the deterministic model showed that a very small proportion of alleles from the wild-type population are rescued through recombination and thus present at the
end of the simulation. However, recombination and selection are stochastic processes and, to make more robust and insightful predictions, a stochastic model of selection and recombination is needed. This approach is developed in the next chapter.
6 Can bacterial recombination disrupt the hitch-hiking effect of a selective sweep?

6.1 Introduction

Surveys of bacterial genetic diversity have revealed many complex and differing patterns and there is currently a vigorous debate about how these should be interpreted (271). A paradoxical aspect of the genetic structure of bacterial populations is that although bacteria can have very large census population sizes, their neutral diversity (the amount of polymorphism found in non-coding regions or that results in synonymous substitutions) is much less than would be expected in populations of this size at a steady state (272). This dearth of neutral diversity has been used as evidence of the impact of selection on bacterial population structure (273). While the reality may reflect the interplay of many factors such as ecological differentiation, metapopulation structure, bacterial recombination, multi-species competition, repeated population bottlenecks and boom-bust dynamics, simple models allow us to focus on identifying the signature and impact of individual factors on bacterial populations (271).

In the previous chapter I reviewed published literature which examined the impact of a selective sweep on genetic diversity. I summarised insights from models, the results of which predict the effect of frequent or absent recombination when an advantageous allele and/or genotype sweeps to fixation. Under the assumptions of the asexual model, as there is no recombination to separate an advantageous mutation from the background on which it arises, the whole genotype will sweep to fixation. This results in a population composed of the descendents of the original genotype carrying the novel advantageous allele, while all other genetic diversity present in the population before the selective sweep will be lost. In contrast, under the assumptions of the sexual model, frequent recombination allows the advantageous allele to be transferred to other genomes and alleles at neutral loci to be transferred to the sweeping genotype. In this model diversity is only lost over a small area of the genotype surrounding the selected locus.

6.1.1 Selective sweeps as a speciation mechanism: the ecotype model

Selective sweeps have been proposed by some authors as the key mechanism regulating genetic diversity in bacterial populations and providing a mechanism of speciation (274,275). Results from the simple asexual model predict that when a novel fitter allele arises at a locus in a bacterial population, this allele will increase in frequency and sweep to fixation in the population. In the
absence of recombination, one consequence of a selective sweep is the extinction of all genetic diversity at all loci other than those within the sweeping genotype. This is referred to as hitch-hiking, as neutral alleles which are associated with the novel fitter allele hitch-hike to fixation with it. This process has been suggested as a mechanism of adaptation and a driver of population divergence in the 'ecotype model' (274). An ecotype is defined as a population of bacteria that are ecologically similar to each other and where genetic diversity is limited within the ecotype by selective sweep events (259). A selective sweep occurs when a mutant arises that is better able to use the resources within the ecotype. The offspring of this mutant grow faster and leave more descendents so that the population eventually consists only of these offspring. As a result genetic variation that existed within other genotypes becomes extinct as they are lost from the population (274).

In the ecotype model bacterial populations diverge into new species when a mutation creates a new phenotype that allows a bacteria and its descendents to live in a new ecological niche. The mutation may allow the new phenotype to utilise an environmental resource that the parent bacteria are unable to. These bacteria are assumed to be protected from selective sweeps in the parent ecotype as this population is ecologically distinct and recombination is assumed to be practically absent between niches (258,276). The two populations can then exist independently and behave as distinct species even though they may exist in the same physical location. Over time, new, fitter bacteria will arise in each population through mutation and sweep to fixation within the separate ecotypes, allowing the two populations to further diverge away from each other. In this way distinct phenotypic clusters exist independently in ecological niches while potentially sharing a physical location and maintaining transient amounts of diversity within a cluster (258).

Cohan has expanded this basic model to include several scenarios, one of which incorporates the migration of strains through human transport thus allowing multiple genotypes within an ecotype (albeit transiently). Another suggests that if advantageous mutants arise frequently, frequent sweeps would not allow diversity to accumulate at all. Further work examines small populations of bacteria where genetic drift is capable of maintaining the same dynamic as selection (274).
Figure 6.1 Stable ecotype model. Ecotype formation events are represented by a change in colour. Periodic selection events are represented by asterisks. Black lines above the phylogenies represent the sequence clusters that should be observed in samples of the present day population. The Stable Ecotype model, in which ecotype formation is relatively rare, and periodic selection events frequently purge all diversity from within each ecotype, therefore predicts a one-to-one correspondence between sequence clusters and ecotypes. Reproduced from (277).

6.1.2 Homologous recombination can prevent a hitch-hiking event occurring

Under the ecotype model recombination is a mechanism which generates novel allelic combinations that may establish a novel phenotype. The Cohen et al. contend that recombination is so infrequent in bacterial populations that they essentially evolve as a population of asexual organisms (274). This contention has been criticised by other authors who believe that recombination would prevent a selective sweep from purging diversity from within an ecotype and would prevent two species successfully diverging. If bacterial populations evolve as sexual populations rather than asexual ones recombination could re-assort alleles at loci in the genome other than the locus under selection. This would result in genetic diversity present in the wild-type population before the selective sweep being retained and the hitch-hiking effect of the selective sweep would be prevented (278). Recombination between ecotypes and across loci other than those that convey phenotypic difference would mean that two ecotypes could be almost identical at all loci other than those under selection. Thus recombination could prevent the formation of new distinct species (279).

6.1.3 Interpreting the outcome of a selective sweep

As the ultimate aim of any evolutionary model is to produce patterns which are a realistic depiction of the outcome of evolutionary processes in nature, the second aim of this chapter is to examine which summary statistics provide an accurate description of the process and outcome of selection.
One method of describing the genetic diversity in bacterial populations is MLST. This is a method of bacterial typing which I described in the first chapter. It is a typing scheme where seven housekeeping genes are sequenced and each different sequence at a locus is assigned a different allele number. The allele numbers at each of the seven loci are used to define the sequence types of each bacterial isolate in the population (101,280,281). A diversity index is a mathematical measure of diversity in a community. In the context of microbial population genetics and MLST it provides a measure of sequence type (ST) diversity.

Diversity indices provide more information about population composition than ST richness (i.e. the number of sequence types present); they also take the relative abundances of different STs into account. Consider two communities of 100 bacteria each and composed of 10 different sequence types. One community has 10 bacteria of each ST, the other has one bacteria of each of nine STs, and 91 bacteria of the tenth ST. Clearly the first population is more diverse, but both populations have the same ST richness. By taking relative abundances into account, a diversity index depends not only on ST richness but also on the evenness, or equitability, with which bacteria are distributed among the different STs.

Two methods of summarising the diversity in a bacterial population can be used to summarise MLST data: allelic distributions (nearest neighbour and mismatch) (106) and Simpson’s \( D \) (292).

### i. Allelic distributions

An allelic distribution is a plot of the distribution of allelic distances between isolates. When using the MLST two isolates can have an identical ST, or differ at one to seven loci. That is, a pair of bacteria which have the same alleles at five of seven loci are two alleles distance apart. The mismatch distribution is the proportion of strains that differ at \( i=0, 1, 2, 3, 4, 5, 6 \) or 7 loci. Defined analytically this distribution can be expressed for the multi locus case of strains which differ at \( i \) loci when \( k > 0 \) by the equation at equilibrium:

\[
F_k^i = \frac{(i-k+1)F_{k-1}^i + \rho (F_{k-2}^i F_{k+1}^i + F_{k-1}^i F_{k+2}^i)}{1 + \rho + (i-k)\theta}
\]

And when \( k = 0 \):

\[
F_0^i = \frac{1 + \rho F_0^{i-1} F_0^1}{1 + t\theta + t\rho}
\]
In equation 6-1 $\rho$ is the population recombination rate and $\theta$ is the population mutation rate. This measure is described in more detail in (106).

The nearest neighbour distribution is a description of the distribution of the proportion of pairs of isolates whose nearest relative differs at 0, 1, 2, 3, 4, 5, 6 or 7 loci from it.

**ii. Simpson’s diversity index ($D$)**

$D$ is a simple mathematical measure that characterizes species diversity, or in this context ST diversity in a population (292). In population genetics $D$ is equivalent to the heterozygosity of a sample, in this context it is the probability that two isolates selected at random from the population will have different STs.

Simpson’s $D$ is defined as:

$$D = \frac{1 - \sum_{i=1}^{m} x_i^2}{1 - 1/N}$$

The fraction with ST $i$ is defined as:

$$x_i = \frac{n_i}{N}$$

The equation can be derived as follows. The probability that the a single bacterium sampled at random will belong to the $i^{th}$ group is $x_i = \frac{n_i}{N}$. The probability that two bacteria sampled consecutively will belong to that group is $x_i^2$. This can then be summed for all the STs (of which there are $m$ in number) to give the probability that two bacteria consecutively sampled will be of the same ST. This summation is subtracted from 1 to give the probability that two consecutively sampled bacteria will have different STs.

The relationship between Simpson’s $D$ and the proportion of bacteria that differ at none of seven loci, i.e. the first point on the mismatch distribution is described by equation 6-4.

$$D = 1 - F_0^7$$

Another aspect of observing and interpreting selective sweeps in nature is selecting a suitable sample of the population to study. Using a mathematical model I can explore the impact of a selective sweep on a population; how long an advantageous allele takes to sweep to fixation, how long it remains at fixation and how the population structure is expected to be disturbed by the
selective sweep. It also allows me to examine the level of population discrimination at which I would need to examine the effect of a sweep.

6.1.4 Aims

In this chapter I will to address the first objection to the ecotype model: that asymmetric bacterial recombination is frequent enough in bacterial populations to prevent a selective sweep acting to remove genetic diversity over the whole genotype within an ecotype. In the previous chapter I showed that as bacterial recombination was a distinct process from its eukaryotic counterpart a specific model was needed to study this process. To this end I will:

- Expand the ‘bacterial’ or asymmetric model outlined in Chapter 5 to include stochastic processes,
- Use a stochastic model that allows recombination and mutation to examine the effect of selection on overall diversity,
- Examine the scenario of clonal interference to ascertain whether the addition of this complexity prevents the loss of genetic diversity as a result of the sweep.

I will also examine how different measures of diversity record the impact of a selective sweep. To this end I will:

- Adapt a stochastic eight locus model of bacterial evolution with recombination and mutation to incorporate a selective sweep,
- Examine the values of Simpson’s D and mismatch distribution at different points in the simulation.
6.2 Stochastic Two Locus Model

In the previous chapter I used a two locus deterministic model to illustrate the difference between bacterial homologous asymmetric recombination and eukaryotic symmetric recombination. The results of this model showed that for most of the parameters examined very small proportions of the alleles from the wild-type population were present in the final population after the selective sweep. The results of the two locus asymmetric deterministic model are illustrated in Figure 6.2. As recombination and selection are stochastic processes, I adapted the two locus model to incorporate stochastic processes into both the selection of bacterial genotypes at each generation and the process of recombination.

![Figure 6.2](image)

**Figure 6.2** Two Locus Asymmetric deterministic model. 
Proportion of the population of recombinant type MX. The population recombination rate $\rho$ is plotted on the log scale against proportion of MX in the final population. The results are shown for selection coefficients $s$ of 0.001 (dotted line), 0.01 (dashes line) and 0.1 (solid line).
6.2.1 Stochastic Model Structure

The population simulated by the stochastic model is the same as that simulated by the deterministic model in Chapter 5. When illustrating this population I considered it as composed of three stages, pre-sweep (the start of the simulation), the intermediate phase (illustrated by demonstrating the recombination and selection events) and the post-sweep population, when the wild-type is extinct and all bacteria carry the advantageous allele, illustrated in Chapter 5, figure 5.1.

As in Chapter 5, each bacterium has two loci, the locus under selection and a neutral locus. In the pre-sweep population there is one bacterium carrying an advantageous allele (M), all other bacteria carry the wild-type allele (W) at the selected locus. At the second neutral locus one bacterium carries a Y allele while all other bacteria carry an X allele. The X allele represents all other wild-type neutral diversity present in the pre-sweep population at this locus rather than one particular allele. At the start of the simulation the M allele is exclusively associated with the Y allele and the W allele is exclusively associated with the X allele. Throughout this chapter I refer to the WX genotype as the wild-type and the MY genotype as the mutant type. Genotypes MX and WY, created through homologous recombination between the wild-type and mutant genotype are referred to as hybrid genotypes.

The model is extended from the Wright-Fisher model of allele frequency change, and incorporates selection and recombination (282). In the Wright-Fisher model offspring are sampled from the parent generation which then dies, so generations are discrete and non-overlapping. The number of different types of bacteria at each generation is simulated under a multinomial distribution dependant on the frequency of the genotypes (WX, WY, MX or MY) in the population from which the next population is drawn. For computational efficiency the multinomial distribution is described as using the binomial distribution commands in C++ so that , the number each genotype that appears in a sample size N

\[ B(n, p) \]

distributed, where the \( p \) is the frequency of the genotype at that timestep. For each subsequent genotype the population is sampled without replacement.

The first step in each simulation of the model defines the proportions of each allele present in the population at time \( t \), \( p'W, p'M, p'X, p'Y \). The equations describing this step are detailed in 6-5. The next step in the programme is the selection event, detailed in equation group 6-6 below. After the selection event the population is defined under the multinomial distribution.
(N'wx, N'wy, N'my, N'mx), his step is detailed in equation group 6-6 and the frequencies of each genotype at this point (f'WX, f'WY, f'MY, f'MX) are defined in equation group 6-7.

6-5

\[
p'W = \frac{N_{wx}(t) + N_{wy}(t)}{N_{wx}(t) + N_{wy}(t) + N_{mx}(t) + N_{my}(t)}
\]

\[
p'M = \frac{N_{mx}(t) + N_{my}(t)}{N_{wx}(t) + N_{wy}(t) + N_{mx}(t) + N_{my}(t)}
\]

\[
p'X = \frac{N_{wx}(t) + N_{mx}(t)}{N_{wx}(t) + N_{wy}(t) + N_{mx}(t) + N_{my}(t)}
\]

\[
p'Y = \frac{N_{my}(t) + N_{wy}(t)}{N_{wx}(t) + N_{wy}(t) + N_{mx}(t) + N_{my}(t)}
\]

6-6

\[
f'WX = \frac{N_{wx}(t)}{(N_{wx}(t) + N_{wy}(t)) + (N_{mx}(t) + N_{my}(t))(1 + s)}
\]

\[
f'WY = \frac{N_{wy}}{(N_{wx}(t) + N_{wy}(t)) + (N_{mx}(t) + N_{my}(t))(1 + s)}
\]

\[
f'MY = \frac{N_{my}(1 + s)}{(N_{wx}(t) + N_{wy}(t)) + (N_{mx}(t) + N_{my}(t))(1 + s)}
\]

\[
f'MX = \frac{N_{mx}(1 + s)}{(N_{wx}(t) + N_{wy}(t)) + (N_{mx}(t) + N_{my}(t))(1 + s)}
\]

6-7

\[N'wx \sim Bin(f'WX, N)\]

\[N'wy \sim Bin\left(\frac{f'WY}{f'WX + f'MX + f'MY}, N - N'wx\right)\]

\[N'mx \sim Bin\left(\frac{f'MX}{f'MX + f'MY}, N - N'wx - N'wy\right)\]

\[N'my \sim Bin\left(\frac{f'MY}{f'MY}, N - N'wx - N'wy - N'mx\right)\]

6-8

\[
f'WX = \frac{N_{wx}}{(N'wx + N'wy) + (N'mx + N'my)}
\]

\[
f'WY = \frac{N_{wy}}{(N'wx + N'wy) + (N'mx + N'my)}
\]

\[
f'MX = \frac{N_{mx}}{(N'wx + N'wy) + (N'mx + N'my)}
\]

\[
f'MY = \frac{N_{my}}{(N'wx + N'wy) + (N'mx + N'my)}
\]
Recombination then occurs in the population through a stochastic three-step process. In stage one (see equation group 6-9), the probability of an allele’s availability to recombine with a recipient genome is simulated under the binomial distribution at rate $r$ (the recombination rate) depending on the frequency of the allele in the population. I derived $r$ in terms of the population recombination rate so that $r = \rho / 2N$, where $\rho$ is the population recombination rate and $N$ is the population size.

Alleles enter the recipient genome simulated under the multinomial distribution dependent on the proportion of free alleles estimated in stage one and the frequency of the relevant genotype in the population detailed in equation groups 6-10, 6-11, 6-12 and 6-13. Stage three simulates the integration of recombinant alleles through calculating the number of different types now present in the population (see equation group 6-14). The new generation $(t+1)$ is then defined as detailed in equation group 6-15.

6-9

\[ pW \sim \text{Bn}(r, N'wx + N'wy) \]

\[ pM \sim \text{Bn}(r, N'mx + N'my) \]

\[ pX \sim \text{Bn}(r, N'mx + N'wx) \]

\[ pY \sim \text{Bn}(r, N'my + N'wy) \]

6-10

\[ qW \rightarrow WX \sim \text{Bn}(\tfrac{f'_{WX}}{f'_{WX} + f'_{MX} + f'_{MY}}, pW - (qW \rightarrow WX)) \]

\[ qW \rightarrow WY \sim \text{Bn}(\tfrac{f'_{WY}}{f'_{WY} + f'_{MX} + f'_{MY}}, pW - (qW \rightarrow WY)) \]

\[ qW \rightarrow MX \sim \text{Bn}(\tfrac{f'_{MX}}{f'_{MX} + f'_{MY}}, pW - (qW \rightarrow WY) - (qW \rightarrow MX)) \]

\[ qW \rightarrow MY \sim pW - (qW \rightarrow WX) - (qW \rightarrow WY) - (qW \rightarrow MX) \]

6-11

\[ qM \rightarrow WX \sim \text{Bn}(\tfrac{f'_{WX}}{f'_{WX} + f'_{MX} + f'_{MY}}, pM - (qM \rightarrow WX)) \]

\[ qM \rightarrow WY \sim \text{Bn}(\tfrac{f'_{WY}}{f'_{WY} + f'_{MX} + f'_{MY}}, pM - (qM \rightarrow WX) - (qM \rightarrow WY)) \]

\[ qM \rightarrow MX \sim \text{Bn}(\tfrac{f'_{MX}}{f'_{MX} + f'_{MY}}, pM - (qM \rightarrow WX) - (qM \rightarrow WY) - (qM \rightarrow MX)) \]

\[ qM \rightarrow MY \sim pM - (qM \rightarrow WX) - (qM \rightarrow WY) - (qM \rightarrow MX) \]

6-12

\[ qX \rightarrow WX \sim \text{Bn}(f'_{WX}, pX) \]
\[ q^X \sim WY, Bn\left(\frac{f^{\prime}WY}{f^{\prime}WY + f^{\prime}MX + f^{\prime}MY}, pX - (qX \rightarrow WX) \right) \]

\[ q^X \sim MX, Bn\left(\frac{f^{\prime}MX}{f^{\prime}MX + f^{\prime}MY}, pX - (qX \rightarrow WX) - (qX \rightarrow WY) \right) \]

\[ q^X \sim MY, pX - (qX \rightarrow WX) - (qX \rightarrow WY) - (qX \rightarrow MX) \]

6.13

\[ q^Y \sim WX, Bn(f^{\prime}WX, pY) \]

\[ q^Y \sim WY, Bn\left(\frac{f^{\prime}WY}{f^{\prime}WY + f^{\prime}MX + f^{\prime}MY}, pY - (qY \rightarrow WX) \right) \]

\[ q^Y \sim MX, Bn\left(\frac{f^{\prime}MX}{f^{\prime}MX + f^{\prime}MY}, pY - (qY \rightarrow WX) - (qY \rightarrow WY) \right) \]

\[ q^Y \sim MY, pY - (qY \rightarrow WX) - (qY \rightarrow WY) - (qY \rightarrow MX) \]

6.14

\[ n^{\prime}WX = nWX - (qM \rightarrow WX) - (qY \rightarrow WX) + (qW \rightarrow MX) + (qX \rightarrow WY) \]

\[ n^{\prime}WY = nWY - (qM \rightarrow WY) - (qX \rightarrow WY) + (qW \rightarrow MY) + (qY \rightarrow WX) \]

\[ n^{\prime}MX = nMX - (qW \rightarrow MX) - (qY \rightarrow MX) + (qM \rightarrow WX) + (qX \rightarrow MY) \]

\[ n^{\prime}MY = nMY - (qW \rightarrow MY) - (qX \rightarrow MY) + (qM \rightarrow WY) + (qY \rightarrow MX) \]

6.15

\[ nWX(t+1) = n^{\prime}WX \]

\[ nWY(t+1) = n^{\prime}WY \]

\[ nMY(t+1) = n^{\prime}MY \]

\[ nMX(t+1) = n^{\prime}MX \]

### 6.2.2 Selection of recombination rates and selection coefficients

The model was simulated for a range of parameter values. The population size N was $10^6$. Values for the population recombination rate $\rho$ were simulated from 1 to 10000 per generation. Values for the selection coefficient $s$ were simulated from $10^{-3}$ to 0.9 so that the fitness of the M allele was varied from 1.001 to 1.9.

In this analysis I used a wide range of recombination rates. These have been estimated previously and are described in detail in (106,283): the authors used the mismatch distribution, maximising the
multinomial log likelihood with respect to the values of $\rho$ (and $\theta$ the population mutation rate). This was fitted to the mismatch distribution calculated from the entire contents of publically available MLST datasets for *Neisseria meningitidis* (http://www.pubmlst.org/neisseria/), *Staphylococcus aureus* (http://www.saureus.mlst.net), *S. pneumoniae* (http://www.spneumoniae.mlst.net), *Streptococcus pyogenes* (http://www.spyogenes.mlst.net/), and *Helicobacter pylori* (http://www.pubmlst.org/helicobacter/) among others (283).

*S. aureus* has been shown to be almost clonal with alleles changing by point mutation about sixteen times more commonly than by recombination (102), with a population recombination rate of 2.4 (283). At the other end of the scale, *H. pylori* is estimated to have a high population recombination rate of 73.3 (283).

In this model I used a wide range of selection coefficient values. Previous laboratory and theoretical studies of the distribution of the fitness effects of advantageous mutants have shown that fitness is distributed exponentially (284); most advantageous mutations have small relative fitness effects. In laboratory studies the relative fitness of detectable advantageous mutations has been shown to range between $\sim 1.05$ and 1.25 when the ancestral population is well adapted to the laboratory environment (284, 285). The range of parameters for the fitness effects of mutations are illustrated in Figure 6.3. In laboratory studies it is difficult to detect small changes in fitness advantage; therefore the lower bound of selection coefficients is higher than that which is likely to occur in nature (285). The choice of the lower bound for $s$ was taken from a review of literature of past models of selective sweeps (286) and also what was computationally feasible to simulate. When $s \leq 0.0009$ no selective sweeps occurred in 1000 simulations of the model, therefore $s = 0.001$ is the lower bound of the selection coefficient in these simulations. The parameter space studied in the model is considered to be a reasonable representation of the possible range of fitness effects in a bacterial population.
Figure 6.3 Distribution of fitness effects.
Distribution of fitness effects among 18 advantageous mutants assayed in the permissive environment. Bars indicate the observed distribution of fitness effects. The dots indicate expected values given an exponential distribution as calculated by the authors. Reproduced from (285).

6.2.3 Model outputs

The impact of recombination on diversity is calculated from three outputs in the two locus model. The first is the proportion of the post sweep population that is made up of the MX genotype. The second is the probability that a genotype carrying an X allele at the neutral locus can sweep to fixation, defined as π. The third is the number of X alleles which become associated with the M allele through recombination over the duration of the simulation.

6.2.4 Results

i. Proportion of X alleles retained

Figure 6.4 shows the average value of P(MX) from multiple sweep events that occurred over 1000 simulations of the model for each value of ρ and s. As can be seen in Figure 6.4 the proportion of MX genotypes is small for a high value of s and low value of ρ. For example, when ρ < 10 and s >
0.001 less than 10% of the final population is composed of MX genotypes. When $\rho$ is very much higher a substantial proportion of X alleles can be rescued, particularly when $s$ is small. Values of $\rho = 1000$ are considered implausibly high. At an upper-bound of $\rho = 200$, which is still well in excess of recombination rates that have been measured, the proportion of MX in the final population varies from 2% for $s = 0.1$ to 60% when $s = 0.001$.

![Graph](image)

**Figure 6.4** Relationship between recombination rate, selection coefficient and proportion of MX in the final population in the two locus stochastic model.

The population recombination rate $\rho$ is plotted on the log scale against proportion of MX in the final population when a sweep has occurred. The results are shown for selection coefficients $s$ of 0.001 (dotted line), 0.01 (dashes line) and 0.1 (solid line). The results are presented for an average of 1000 simulations of the model.

**ii. Impact of recombination on the hitchhiking event**

When a selective sweep occurs in a clonal population the advantageous allele sweeps to fixation along with the rest of the genotype on which the advantageous allele arose. At the end of the sweep all alleles in the population will be alleles or descendents of alleles present in the original bacterium carrying the advantageous allele. In a recombining population (even at rates seen in bacterial populations), Doolittle and others predict that when the mutant allele comes to fixation, alleles at other loci will often be those rescued from the wild-type bacteria in the pre-sweep population (287,279).

To assess the probability of this prediction I created a metric $\pi$, which is the proportion of selective sweeps that result in the fixation of an X allele at the neutral locus. To put this in terms of the
components of the model, if in 5 of 100 sweeps, P(MX)=1 at the last generation of the simulation, then \( \pi = 5/100 \).

The lower bound of \( \pi \) was set to 0.001 when fixation of an X allele had failed to occur within 1000 sweep events. Below this boundary the fixation probability was considered to be too low to be computationally feasible to measure. In my simulations the fixation of the X allele did not occur in 1000 selective sweeps when \( \rho \leq 100 \), even when \( s \) was at the lower bound of 0.001, as shown in Table 6.1.

A \( \rho \) of 100 is a very high population recombination rate; for example \( \rho \) of 73.4 has been estimated from populations of \( H. pylori \) (283). The result of these simulations demonstrated that the complete disruption of the purging effect of the selective sweep is an extremely rare event in populations with low, intermediate or high population recombination rates. Other than populations such as \( H. pylori \) with very high recombination rates, a selective sweep at a selected locus will result in the loss of alleles that were associated with the wild-type population at other loci.

### Table 6.1 Probability of fixation of neutral allele in 1000 selective sweeps as a function of recombination and selection.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( s )</th>
<th>0.001</th>
<th>0.01</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.20</td>
<td>0.01</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
</tbody>
</table>

iii. **Impact of recombination over time**

In the analysis of the model simulations I have shown that small but detectable amounts of the X allele (which in the wild population was exclusively associated with the wild-type genotype WX) remain in the population at relatively high recombination rates when the selection coefficient is very small (\( s \leq 0.001 \)). However, I have also shown that even at high recombination rates when \( \rho \geq 100 \), the probability of the X allele becoming fixed in the population (to the exclusion of the Y allele) is very rare. Thus, when the X allele is maintained in the population it exists with the Y allele in the population also.
In order to examine further the relationship between $\rho$, $s$ and the impact of the selective sweep on descent and why so few hybrids are observed in populations that recombine at high rates, I plotted the creation of new hybrids against time along with the prevalence of the wild-type population. Figure 6.5 shows the number of X alleles incorporated into a genotype carrying the advantageous mutant allele over time. In this simulation the total population size is $10^6$ and $\rho$ is 100. The number of X alleles incorporated reaches a maximum when the WX and MY genotypes are at intermediate levels. The number of generations until the extinction of the WX genotype increases as the selective advantage of the selected mutant allele decreases. The longer bacteria with a WX genotype are present at intermediate levels in the population, the more likely a recombination event will occur between the WX genotype and the MY genotype. At the start of the simulation the creation of the hybrid is unlikely (unless recombination is at a very high rate) as the MY genotype is extremely rare. At the end of the simulation the opposite is true: the WX genotype is extremely rare so recombination with the MY genotype is an extremely rare event.

The amount of X alleles in the final population is dependent on the amount of time the wild-type genotype and mutant genotype are at intermediate values, as well as the recombination rate. This period of time is dependent on how fast the sweep is, the speed of which is dependent on the relative fitness of the advantageous allele, which in turn is determined by the selection coefficient. As the value of $s$ decreases, the amount of time the WX genotype remains in the population increases and the amount of the X alleles that remain in the population increases.
Figure 6.5 Impact of recombination over time. 
A) Number of pre-sweep X alleles from the wild-type WX genotype rescued by becoming associated with the M through recombination over time. B) Proportion of the wild-type genotype, WX, in the population over time. The results are shown for selection coefficients $s$ of 0.1 (grey), 0.01 (blue), 0.001 (red). $p = 100$ and $N=106$.

As a result, the lower the value of $s$ the greater the amount of wild-type alleles present at neutral loci in the population. I examined the proportion of MX genotypes in the final population for very low selection coefficients and how probable a selective sweep was for a given selection coefficient. To do this I used an estimation of the fixation probability of an allele taken from the published literature on nearly neutral theory, which is an extension of the neutral theory. Under the neutral theory a neutral allele has a $1/N$ probability of becoming fixed in the population. The nearly neutral theory expanded the neutral theory in response to the observation that the fitness of a mutation is not neutral or non-neutral but may be a range of values (288-290). As the selection coefficient approaches zero, the relative fitness of the mutant allele approaches neutrality. The nearly neutral theory states that the effect of a selection coefficient can be considered neutral when $sN_e > 1$ where $N_e$ is the effective population size. The probability of an allele with a selection coefficient $s$ becoming fixed in the population is shown by equation 6-16 where $N$ is the population size, $s$ is the selection coefficient and $p$ is the initial frequency of the mutant.

$$u = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}$$
Figure 6.6 Relationship between the fixation probability of the M allele and P(MX), the proportion of MX in the post-sweep population (Y axis).

\[ u \equiv \frac{1 - e^{-2Nsp}}{1 - e^{-2Ne}} \]  

The fixation probability is calculated from the equation derived by Ohta et al. (288), where \( N = 106 \) is the population size, \( s \) is the selection coefficient and \( p \) is the initial frequency of the mutant.

In Figure 6.6, when \( u \) is \( 2 \times 10^{-5} \), the selection coefficient \( s \) is \( 10^{-5} \) and when a sweep occurs in the population, 100% of the population is composed of MX genotypes. However, with these parameter values only one selective sweep occurred out of 5000 simulations of the stochastic model. With a selection coefficient of \( 10^{-6} \) no sweeps occur out of 5000 simulations. Thus, although a selection coefficient may be considered non-neutral, a sweep may in fact be very rare, so rare that it makes it computationally unfeasible to estimate.
6.3 Eight Locus Model of a Selective Sweep in a Mutating and Recombining Population

In the models described in Chapter 5 and section 6.2, I examined whether recombination can disrupt the hitch-hiking effect of a selective sweep. I showed that the majority of genetic diversity present in wild-type genotypes in the pre-sweep population is lost as the result of a selective sweep. This analysis was only concerned with the fate of the neutral wild-type alleles in the population. The question of interest was which genotype in the pre-sweep population neutral alleles in the post-sweep population were related to: the wild-type genome or the genotype carrying the advantageous allele. As the wild-type alleles could only be ‘saved’ through recombination the effect of mutation was not considered in the model.

At the beginning of this chapter I outlined that the ultimate aim of any model is to provide outputs which can inform the analysis of natural populations of bacteria. To understand the impact of a selective sweep on a natural population of bacteria I sought to use a more realistic model that incorporated mutation. In addition I aimed to format the model so that the results of simulations could inform analysis of datasets collected from natural bacterial populations. These populations have commonly been typed using MLST, and so any output from the model should be compatible with this format. To fulfil this aim I used a model that incorporates mutation and recombination at multiple loci.

6.3.1 Model structure

The model builds on earlier work which allowed simulations of bacterial populations that can be compared to experimental data derived using MLST (106,283). This original model was an extension of the neutral infinite alleles model (291) describing neutral variation at seven loci evolving by mutation and recombination. A simple modification of this is a parsimonious model for describing population genetic variation in S. pneumoniae (106). It is parameterised by population mutation and recombination rates, denoted θ and ρ respectively, and by population size N. θ = 2mN, with m the mutation rate and ρ = 2τN with τ the recombination rate.

The model illustrated in Figure 6.7 considers N individuals with recombination and mutation occurring at seven neutral loci but includes an eighth locus on which selection acts. The model is parameterised by rates θ and ρ, as above, together with the fitness advantage of the advantageous mutant. The initial population consists of a single genotype. At each generation of the model,
genotypes are selected at random from the previous generation, then these genotypes can mutate, recombine or both events can occur. At each generation recombination can also occur at the eighth loci. The probability of recombination occurring at each generation to each individual genotype is simulated through random number generation with probability of $r$ the recombination rate. The donor allele is randomly chosen from other genotypes in the population and recombination always changes a whole locus. Mutation can occur at rate $m$ throughout the simulation at the seven neutral loci and always results in a novel allele being generated. The population diversifies through mutation and recombination until it is at an equilibrium level of diversity.

When the population is at equilibrium one genotype is selected at random and a mutation occurs at the eighth locus generating the advantageous allele. This allele conveys a fitness of $1+s$ to any genotype in which it occurs. If the advantageous allele is lost through recombination then the fitness advantage is lost. If the advantageous allele is acquired through recombination then the genotype gains the associated fitness advantage.
Figure 6.7 Eight locus individual model of selection. This illustration shows the model for a population of five individuals. The bacterial strain at each generation (time interval t) is defined by seven neutral loci and one locus which is under selection. At each generation every individual can infect another. Mutation (red line) and recombination (green line) can occur at each transmission step. Mutation occurs at rate m and recombination occurs at rate r. Recombination and mutation can affect more than one allele in a single step. Each mutation creates a new allele. A recombination event results in an individual bacterium inheriting an allele from a random donor. Once dynamic equilibrium has been achieved a mutation at the eighth locus is assigned the fitness 1+s, this mutation is shown here at 1*. Any genotype that carries this mutation has the fitness 1+s. Illustration adapted from (106).
iv. **Model outputs**

The overall diversity of the simulated population is measured by Simpson’s $D$, which is a measure of bacterial strain (genotypic) diversity in a population (292). By taking strain relative abundance into account it is possible to measure not only the strain richness but also the distribution of those strains in the population. Simpson’s $D$ is defined as in equations 6-2 and 6-3.

This model is a multilocus simulation incorporating recombination and mutation. Simpson’s $D$ measures overall diversity, which is the rate at which new strains are generated either by recombination or mutation. Even when a measure of strain diversity is identical for two populations, if the rate of recombination relative to mutation differs in the two populations then the population structure of those populations could differ. This is measured using the allelic mismatch distribution.

v. **Parameter selection**

The recombination rate affects the diversity of the population, even if selection is not present. Hanage et al. explored the effect of recombination on the population structure in a population evolving under the neutral model (283), at a fixed heterozygosity of $D=0.88$. Using the allelic mismatch distribution, the authors showed that increasing $r/m$, while keeping the rate at which new strains are generated constant, impacts the population structure, even though the diversity of strains in the population may be constant, illustrated in Figure 6.8. How those strains are related i.e. how different they are from each other can vary markedly due to recombination.

![Figure 6.8](image.png)

Figure 6.8 Increasing recombination in a population with constant homozygosity.

The allelic mismatch distributions $\bar{F}$ are shown as solid bars. Three combinations of $\theta$ and $\rho$ are shown along an isocline of $D = 0.88$: a clonal population where $\theta=7$ and $\rho=0$, a population with intermediate levels of recombination ($\rho/\theta=0.5$, $\theta=4.75$, $\rho=2.38$), and a population with high levels of recombination ($\rho/\theta=5$, $\theta=1.69$, $\rho=8.45$). These simulations had a population size of 2000. Reproduced from (283).
In order to examine the impact of increasing the ratio of recombination to mutation during a sweep the parameters of the population recombination rate and mutation rate were set so that $D$ remains constant under neutrality, as described previously (283). From Figure 6.9 I selected an isocline of diversity where Simpson’s $D = 0.88$, which dictated the recombination and mutation values which allowed this measure of diversity to be constant under neutrality.

In the clonal simulation $\rho=0$ and $\theta=7$, in the intermediate simulation $\rho=2.38$ and $\theta=4.75$ ($\rho/\theta = 0.5$) and in the high recombination simulation $\rho=8.45$ and $\theta=1.69$ ($\rho/\theta = 5$). In all simulations $N=2000$ and the selection coefficient is 0.9, as I sought to examine multiple sweeps (a high selection coefficient was necessary to allow multiple sweeps to be computationally practical). The simulations were run for 1000 generations, until the population reached a dynamic equilibrium and Simpson’s $D = 0.88$. Once dynamic equilibrium was achieved, a mutation at the eighth locus was assigned the fitness $1+s$. The model was simulated for another 1000 generations until the population has returned to equilibrium, which was measured by the value of $D$ returning to 0.88.

Figure 6.9 Effect of mutation and recombination on strain diversity. The predicted strain diversity within the underlying neutral population, measured by the homozygosity source, is plotted as a function of the population mutation rate $\theta$ and recombination rate $\rho$ (lower values of 1–$D$ indicate more diversity). The line $\rho/\theta=1$ is shown. The positions of Streptococcus pneumoniae (S.pn), Neisseria meningitidis (N.m) Staphylococcus aureus (S.a), Helicobacter pylori (H.p), Streptococcus pyogenes (S.py), Burkholderia pseudomallei (B.p) and Bacillus cereus (B.c) based on parameters estimated from all data for those species in the corresponding MLST databases, are shown. Reproduced from (283).

6.3.2 Results

The results of simulations under clonal (A), medium recombination (B) and high recombination (C) parameter sets are shown in Figure 6.10. Each graph shows $D$ measured at each generation for five selective sweeps, each with a baseline measure of $D$ of 0.88. In each simulation the advantageous allele, which has a relative fitness of 1.9, is introduced into the population after 1000 generations.
I sought to examine whether changing the ratio of recombination to mutation affected the depth and the length of the sweep. I therefore simulated within a parameter space which has a consistent measure of $D$ under neutrality for differing ratios of mutation to recombination. Diversity is restored in the population through the generation of novel alleles by mutation and the re-assortment of these novel alleles through recombination. In Figure 6.10, in each simulation the genotype carrying the advantageous mutant sweeps to fixation in the population over a small number of generations, rapidly decreasing the value of $D$. It is evident that the genomic diversity in the population is severely reduced for a short period of time as a result of the selective sweep.
Figure 6.10 Strain diversity over time.
A measure of heterozygosity D is plotted over time for clonal (A), intermediate (B) and high (C) recombination to mutation ratios. Each graph shows the results of 5 selective sweeps. The parameters for ρ and θ are taken from an isocline of D=0.88. In A ρ=0 and θ=7, (ρ/θ = 0) in B ρ=2.38 and θ=4.75 (ρ/θ = 0.5) and C ρ=8.45 and θ=1.69 (ρ/θ = 5). In all simulations N=2000.
In results from clonal simulations shown in Figure 6.10(A), the selective sweep peaks at a lower average value of $D$ than the simulations where $\rho/\theta = 0.5$ or $\rho/\theta = 5$. In simulations where $\rho/\theta = 5$ the population takes longer to return to dynamic equilibrium than in simulations where $\rho/\theta = 0.5$ or where $\rho = 0$.

This apparent effect of recombination is due to the parameterisation of the model. The impact of maintaining constant $D$ is that as the recombination rate increases, the mutation rate must decrease. Thus, at high recombination rates shown in Figure 6.10(C) the mutation rate is low and consequently the population takes longer to return to $D = 0.88$ than the clonal simulation, as novel alleles are generated at a much slower rate. Conversely in a clonal population the mutation rate is high therefore mutations remove the advantageous mutation and decrease the impact of the selective sweep. Thus, increasing the rate of recombination relative to mutation does not have any discernable impact on the population if changes in the population are measured by Simpson’s $D$.

A simple diversity measure, such as $D$, does not reveal the full impact of recombination on a selective sweep, although it provides an excellent measure of the impact of the selective sweep on overall diversity. I therefore looked at the mismatch distribution to observe the effect of a sweep on population structure.

In the results from a simulation of a clonal population, shown in Figure 6.11(A) when the simulation is at equilibrium the proportion of isolates that differ at 0, 1, 2, 3, 4, 5, 6 and 7 loci increase monotonically. The proportion of isolates that differ at all loci will increase as more loci are examined, as under clonality the only way two bacteria can share an allele is through descent. As the simulation continues, each clonal lineage accumulates novel alleles which are only found in its descendents. When selection occurs (shows as black bars), the standing diversity is decreased in the population within 25 generations. At the point of lowest diversity in the population 66% of the population are identical at all loci, while 32% are identical at all but one locus. In these simulations the population is reduced to one clonal lineage with related single locus variants. At 1025 generations there were 125 strains in a population of 2000 bacteria and when equilibrium was re-established the number of strains in the population was 241 and 61% of those strains were discordant over 7 loci.

When a population has a medium recombination to mutation ratio ($\rho/\theta = 0.5$), at equilibrium the proportion of bacteria that are identical at all loci is slightly less than a population with no recombination. However, there is a difference in the proportion of bacteria that differ at five or more loci. Compared to the clonal population, (shown in Figure 6.11(B)) fewer isolates differ at five,
six and seven loci at equilibrium. At the peak of the selective sweep, 40% of the population is alike at all loci while 32% of the population differs at more than six loci. Unlike the clonal population the proportion of isolates that differed at six loci is the same at the peak of the selective sweep and when the population is at equilibrium. At the peak of the selective sweep the population is reduced to two clonal lineages: the bacteria carrying the advantageous mutant and its single locus variants (60% of the population) and a group of bacteria (35%) that differ at six loci. Again, as seen in the clonal model, this is resolved to neutral expectation by 2000 generations.

In the population that has a high ratio of recombination to mutation when the population is at equilibrium most bacteria are identical at three loci. This is due to the fact that in order to maintain constant homozygosity the mutation rate is decreased as the recombination rate increases, therefore the number of novel alleles and strains generated is very low. At the peak of the selective sweep most bacteria (80%) are alike at all loci, the highest level of homozygosity in the three parameter spaces.
Figure 6.11 Mismatch distribution results from eight locus model.
The proportion of bacteria that are discordant at 0, 1...7 loci in clonal (A), intermediate (B) and high (C) recombination to mutation ratios. The parameters for $\rho$ and $\theta$ are taken from an isocline of $D=0.88$. In (A) $\rho=0$ and $\theta=7$, in (B) $\rho=2.38$ and $\theta=4.75$ and (C) $\rho=8.45$ and $\theta=1.69$. In all simulations $N=2000$. The white column is the allelic distribution when $D$ is at the minimum value, the black column is from the population when $D$ has returned to 0.88.
6.4 Clonal interference

Having considered the simplest model of a selective sweep, a two locus model with a single advantageous mutant, and a multilocus model of a selective sweep, I sought to consider a more complex evolutionary scenario. I sought to examine whether a more complex model allowed more of the neutral wild-type alleles to be ‘rescued’ from the pre-sweep population.

In the conventional model of asexual evolution advantageous mutations rarely arise in the population. When they do, the genotype on which they arose expands clonally through the population until it is fixed in the population, purging the population of genetic diversity. This is the strong selection with weak mutation model of asexual evolution and the scenario I modelled in the previous three models studied (286,293,294). When a population is large and has a high mutation rate, advantageous mutations may be common enough that they occur at the same time in the population. In an asexual population, the absence of recombination results in competition between the clones that carry the different advantageous mutations. This phenomenon is known as clonal interference (295). Theoretical models of clonal interference lead to increased fitness effects and longer fixation times of the winning advantageous mutation (296,297,295). Evidence for the occurrence of clonal interference has been inferred from laboratory studies of bacterial populations (239,298-300). Previous studies have examined the fate of the advantageous mutation, but did not address the effect of the clonal interference on genotype wide diversity. In order to examine the effect of two competing advantageous mutations on the standing diversity during a sweep I expanded the original selective sweep model.

6.4.1 Model structure

I extended the two locus selective sweep model to include a second advantageous allele. These two alleles are at the same locus: allele Ma is one advantageous allele and Mb is the competing advantageous allele. As in the original model, W is the wild-type allele at the same locus. The three neutral alleles at the second locus are X, Y and Z. The fitness of W is 1, the fitness of Ma is 1 + sa and the fitness of Mb is 1+sb, where sa and sb are the selection coefficients of the advantageous and interfering mutations respectively and sa > sb. At the beginning of the simulation (pre-sweep population) there is one bacterium carrying the Ma and Y alleles (genotype MaY) and one bacterium carrying the Mb and Z alleles (genotype MbZ). The remainder of the population is the genotype WX.

As shown in Figure 6.12, in the pre-sweep population the X allele is exclusively associated with the W allele (genotype WX), the Y allele is exclusively associated with Ma, and the Z allele is exclusively
associated with Mb. Recombination allows nine possible combinations of alleles, WX, WY, WZ, MaX, MaY, MaZ, MbX, MbY, MbZ.

The model is extended from the Wright-Fisher model of allele frequency change, and incorporates selection and recombination (282). The number of different types of bacteria at each generation is simulated under a multinomial distribution dependant on the frequency of the genotypes (WX, WY, WZ, MaX, MaY, MaZ, MbX, MbY and MbZ) in the population from which the next population is drawn.

The first step in each simulation of the model defines the proportions of each allele present in the population at time \( t \). The equations describing this step are detailed in equation group 6-17. The next step in the programme is the selection event, detailed in equation group 6-18. When selection acts on the population, those bacteria that carry the Ma allele are multiplied by \((1+s_a)\) and those that carry the Mb allele are multiplied by \((1+s_b)\). Those bacteria that do not carry a advantageous Ma or Mb allele are multiplied by 1. After the selection event the population is defined under the multinominal distribution. This step is detailed in equation group 6-19 and the frequencies of each genotype at this point are defined in equation group 6-20.
Figure 6.12 Structure of clonal interference extension.
The initial pre-sweep population is composed of three genotypes, pure wild-type WX (blue), pure mutant type MaY (red) and pure interfering type MbZ (purple). Recombination can occur at rate $\frac{p}{2N}$ between the loci in either direction. Selection increases the proportion of bacteria with the Ma or Mb beneficial mutations according to their selection coefficients, $s_a$ and $s_b$). The fitness of hybrid genotypes are dependent on the allele at the selection locus. In the final population post-sweep population all genotypes contain the Ma or Mb allele (red or purple) at the selection locus and the rescue of pre-sweep X alleles (blue) at the neutral locus is measured by the proportion of the hybrid genotypes MbX and MaX.
\[ p'w = \frac{Nwx(t) + Nwy(t) + Nwz(t)}{Nwx(t) + Nwy(t) + Nwz(t) + Nmax(t) + Nmay(t) + Nmax(z) + Nmbx(t) + Nmby(t) + Nmbz(t)} \]
\[ p'ma = \frac{Nmax(t) + Nmay(t) + Nmax(z)}{Nwx(t) + Nwy(t) + Nwz(t) + Nmax(t) + Nmay(t) + Nmax(z) + Nmbx(t) + Nmby(t) + Nmbz(t)} \]
\[ p'mb = \frac{Nmbx(t) + Nmby(t) + Nmbz(t)}{Nwx(t) + Nwy(t) + Nwz(t) + Nmax(t) + Nmay(t) + Nmax(z) + Nmbx(t) + Nmby(t) + Nmbz(t)} \]
\[ p'x = \frac{Nmax(t) + Nmbx(t)}{Nwx(t) + Nwy(t) + Nwz(t) + Nmax(t) + Nmay(t) + Nmax(z) + Nmbx(t) + Nmby(t) + Nmbz(t)} \]
\[ p'y = \frac{Nmay(t) + Nmby(t)}{Nwx(t) + Nwy(t) + Nwz(t) + Nmax(t) + Nmay(t) + Nmax(z) + Nmbx(t) + Nmby(t) + Nmbz(t)} \]
\[ p'z = \frac{Nmax(z) + Nwz(t) + Nmbz(t)}{Nwx(t) + Nwy(t) + Nwz(t) + Nmax(t) + Nmay(t) + Nmax(z) + Nmbx(t) + Nmby(t) + Nmbz(t)} \]

\[ fwx = \frac{Nwx(t)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ fwy = \frac{Nwy(t)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ fwx = \frac{Nwz(t)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ ma = \frac{Nmax(t)(1 + sa)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ may = \frac{Nmay(t)(1 + sa)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ ma = \frac{Nmax(z)(1 + sa)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ mbx = \frac{Nmbx(t)(1 + sb)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ mby = \frac{Nmby(t)(1 + sb)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]
\[ mbz = \frac{Nmbz(t)(1 + sb)}{(Nwx(t) + Nwy(t) + Nwz(t)) + (Nmax(t) + Nmay(t) + Nmax(z))(1 + sa) + (Nmbx(t) + Nmby(t) + Nmbz(t))(1 + sb)} \]

\[ N'wx \sim Bn(fwx,N) \]
\[ N'wy \sim Bn(\frac{fwy}{fwy + fwx + fmax + may + fmx + mbx + fmbx + fmbx + mbz}N - N'wx) \]
\[ N'wz \sim Bnl\left( \frac{fWZ}{fWZ + fMaX + fMaY + fMaZ + fMbX + fMbY + fMbZ}, N - N'wx - N'wy \right) \]

\[ N'max \sim Bnl\left( \frac{fMaX}{fMaX + fMaY + fMaZ + fMbX + fMbY + fMbZ}, N - N'wx - N'wy - N'wz \right) \]

\[ N'may \sim Bnl\left( \frac{fMaY}{fMaY + fMaZ + fMbX + fMbY + fMbZ}, N - N'wx - N'wy - N'wz - N'max \right) \]

\[ N'maz \sim Bnl\left( \frac{fMaZ}{fMaZ + fMbX + fMbY + fMbZ}, N - N'wx - N'wy - N'wz - N'max - N'may \right) \]

\[ N'mbx \sim Bnl\left( \frac{fMbX}{fMbX + fMbY + fMbZ}, N - N'wx - N'wy - N'wz - N'max - N'may - N'maz \right) \]

\[ N'mby \sim Bnl\left( \frac{fMbY}{fMbY + fMbZ}, N - N'wx - N'wy - N'wz - N'max - N'may - N'maz - N'mbx \right) \]

\[ N'mbz \sim Bnl\left( \frac{fMbY}{fMbY}, N - N'wx - N'wy - N'wz - N'max - N'may - N'maz - N'mbx - N'mby \right) \]

\[ f'WX = \frac{N'wx}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]

\[ f'WY = \frac{N'wy}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]

\[ f'WZ = \frac{N'wz}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]

\[ f'MaX = \frac{N'max}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]

\[ f'MaY = \frac{N'may}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]

\[ f'MaZ = \frac{N'maz}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]

\[ f'MbX = \frac{N'mbx}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]

\[ f'MbY = \frac{N'mby}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz} \]
Recombination then occurs in the population through a stochastic three-step process. In stage one (see equation group 6-21), the probability of an allele’s availability to recombine with a recipient genotype is simulated under the binomial distribution at rate $r$ (the recombination rate) depending on the frequency of the allele in the population. $r$ is derived in terms of the population recombination rate so that $r = \rho/2N$ where $\rho$ the population recombination rate and $N$ is the population size. Alleles enter the recipient genotype simulated under the multinomial distribution, which is dependent on the proportion of free alleles (estimated in stage one) and the frequency of the relevant genotype in the population. This process is detailed in equation groups 6-22 to 6-25. Stage three simulates the integration of recombinant alleles, calculating of the number of different genotypes now present in the population; presented in equation group 6-26. The new generation $(t+1)$ is then defined as detailed in equation group 6-29.

$$f'MbZ = \frac{N'mbz}{N'wx + N'wy + N'wz + N'max + N'may + N'maz + N'mbx + N'mby + N'mbz}$$

Stage three simulates the integration of recombinant alleles, calculating of the number of different genotypes now present in the population; presented in equation group 6-26. The new generation $(t+1)$ is then defined as detailed in equation group 6-29.

6-21

$pW \sim Bnl(r, N'wx + N'wy + N'wz)$

$pMa \sim Bnl(r, N'max + N'may + N'maz)$

$pMb \sim Bnl(r, N'mbx + N'mby + N'mbz)$

$pX \sim Bnl(r, N'max + N'wx + N'mbx)$

$pY \sim Bnl(r, N'may + N'wy + N'mby)$

$pZ \sim Bnl(r, N'maz + N'wz + N'mbz)$

6-22

$qW \rightarrow WX \sim Bnl(f'WX, pW)$

$qW \rightarrow WY \sim Bnl(\frac{f'WY}{f'WY + f'WZ + f'MaX + f'MaY + f'MaZ + f'Mbx + f'Mby + f'Mbz}, pW - (qW \rightarrow WX))$
\begin{align*}
qW \rightarrow WZ & \sim \text{Bnl} \left( \frac{f^\prime WZ}{f^\prime WZ + f^\prime M_a X + f^\prime M_a Y + f^\prime M_a Z + f^\prime M_b X + f^\prime M_b Y + f^\prime M_b Z}, pW \right) \\
& \quad - (qW \rightarrow WX) - (qW \rightarrow WY) \\
qW \rightarrow M_a X & \sim \text{Bnl} \left( \frac{f^\prime M_a X}{f^\prime M_a X + f^\prime M_a Y + f^\prime M_a Z + f^\prime M_b X + f^\prime M_b Y + f^\prime M_b Z}, pW - (qW \rightarrow WX) \right) \\
& \quad - (qW \rightarrow WY) - (qW \rightarrow WZ) \\
qW \rightarrow M_a Y & \sim \text{Bnl} \left( \frac{f^\prime M_a Y}{f^\prime M_a Y + f^\prime M_a Z + f^\prime M_b X + f^\prime M_b Y + f^\prime M_b Z}, pW - (qW \rightarrow WX) \right) \\
& \quad - (qW \rightarrow WY) - (qW \rightarrow WZ) - (qW \rightarrow M_a X) \\
qW \rightarrow M_a Z & \sim \text{Bnl} \left( \frac{f^\prime M_a Z}{f^\prime M_a Z + f^\prime M_b X + f^\prime M_b Y + f^\prime M_b Z}, pW - (qW \rightarrow WX) - (qW \rightarrow WY) \right) \\
& \quad - (qW \rightarrow WZ) - (qW \rightarrow M_a X) - (qW \rightarrow M_a Y) \\
qW \rightarrow M_b X & \sim \text{Bnl} \left( \frac{f^\prime M_b X}{f^\prime M_b X + f^\prime M_b Y + f^\prime M_b Z}, pW - (qW \rightarrow WX) - (qW \rightarrow WY) -(qW \rightarrow WZ) \right) \\
& \quad - (qW \rightarrow M_a X) - (qW \rightarrow M_a Y) - (qW \rightarrow M_a Z) \\
qW \rightarrow M_b Y & \sim \text{Bnl} \left( \frac{f^\prime M_b Y}{f^\prime M_b Y + f^\prime M_b Z}, pW - (qW \rightarrow WX) - (qW \rightarrow WY) -(qW \rightarrow WZ) \right) \\
& \quad - (qW \rightarrow M_a X) - (qW \rightarrow M_a Y) - (qW \rightarrow M_a Z) - (qW \rightarrow M_b X) \\
qW \rightarrow M_a Y & \sim \text{Bnl} \left( f^\prime WY, pW - (qW \rightarrow WX) - (qW \rightarrow WY) -(qW \rightarrow WZ) \right) \\
& \quad - (qW \rightarrow M_a X) - (qW \rightarrow M_a Y) - (qW \rightarrow M_a Z) - (qW \rightarrow M_b X) - (qW \rightarrow M_b Z) \\
\end{align*}

6-23

\begin{align*}
qX \rightarrow WX & \sim \text{Bnl}(f^\prime WX, pX) \\
qX \rightarrow WY & \sim \text{Bnl}(f^\prime WY + f^\prime WZ + f^\prime M_a X + f^\prime M_a Y + f^\prime M_a Z + f^\prime M_b X + f^\prime M_b Y + f^\prime M_b Z, pX) \\
& \quad - (qX \rightarrow WX))
\end{align*}
\[ qX \rightarrow WZ \sim Bnl \left( \frac{f' \, WZ}{f' \, WZ + f' \, MaX + f' \, MaY + f' \, MaZ + f' \, MbX + f' \, MbY + f' \, MbZ}, pX \right. \]
\[ \left. - (qX \rightarrow WX) - (qX \rightarrow WY) \right) \]
\[ qX \rightarrow MaX \sim Bnl \left( \frac{f' \, MaX}{f' \, MaX + f' \, MaY + f' \, MaZ + f' \, MbX + f' \, MbY + f' \, MbZ}, pX - (qX \rightarrow WX) \right. \]
\[ \left. - (qX \rightarrow WY) - (qX \rightarrow WZ) \right) \]
\[ qX \rightarrow MaY \sim Bnl \left( \frac{f' \, MaY}{f' \, MaY + f' \, MaZ + f' \, MbX + f' \, MbY + f' \, MbZ}, pX - (qX \rightarrow WX) - (qX \rightarrow WY) \right. \]
\[ \left. - (qX \rightarrow WZ) - (qX \rightarrow MaX) \right) \]
\[ qX \rightarrow MaZ \sim Bnl \left( \frac{f' \, MaZ}{f' \, MaZ + f' \, MbX + f' \, MbY + f' \, MbZ}, pX - (qX \rightarrow WX) - (qX \rightarrow WY) \right. \]
\[ \left. - (qX \rightarrow WZ) - (qX \rightarrow MaX) - (qX \rightarrow MaY) \right) \]
\[ qX \rightarrow MbX \sim Bnl \left( \frac{f' \, MbX}{f' \, MbX + f' \, MbY + f' \, MbZ}, pX - (qX \rightarrow WX) - (qX \rightarrow WY) - (qX \rightarrow WZ) \right. \]
\[ \left. - (qX \rightarrow MaX) - (qX \rightarrow MaY) - (qX \rightarrow MaZ) \right) \]
\[ qX \rightarrow MbY \sim Bnl \left( \frac{f' \, MbY}{f' \, MbY + f' \, MbZ}, pX - (qX \rightarrow WX) - (qX \rightarrow WY) - (qX \rightarrow WZ) \right. \]
\[ \left. - (qX \rightarrow MaX) - (qX \rightarrow MaY) - (qX \rightarrow MaZ) - (qX \rightarrow MbX) \right) \]
\[ qX \rightarrow MY \sim pX - (qX \rightarrow WX) - (qX \rightarrow WY) - (qX \rightarrow WZ) - (qX \rightarrow MaX) - (qX \rightarrow MaY) \]
\[ - (qX \rightarrow MaZ) - (qX \rightarrow MbX) - (qX \rightarrow MbZ) \]

6-24

\[ qY \rightarrow WX \sim Bnl (f' \, WX, pY) \]
\[ qY \rightarrow WY \sim Bnl \left( \frac{f' \, WY}{f' \, WY + f' \, WZ + f' \, MaX + f' \, MaY + f' \, MaZ + f' \, MbX + f' \, MbY + f' \, MbZ}, pY \right. \]
\[ \left. - (qY \rightarrow WX) \right) \]
\[ qY \rightarrow WZ \sim Bnl \left( \frac{f'WZ}{f'WZ + f'Ma'X + f'Ma'Y + f'Ma'Z + f'Mb'X + f'Mb'Y + f'Mb'Z}, pY \right) - (qY \rightarrow WX) - (qY \rightarrow WY) \]

\[ qY \rightarrow Ma'X \sim Bnl \left( \frac{f'Ma'X}{f'Ma'X + f'Ma'Y + f'Ma'Z + f'Mb'X + f'Mb'Y + f'Mb'Z}, pY - (qY \rightarrow WX) \right) - (qY \rightarrow WY) - (qY \rightarrow WZ) \]

\[ qY \rightarrow Ma'Y \sim Bnl \left( \frac{f'Ma'Y}{f'Ma'Y + f'Ma'Z + f'Mb'X + f'Mb'Y + f'Mb'Z}, pY - (qY \rightarrow WX) - (qY \rightarrow WY) \right) - (qY \rightarrow WZ) - (qY \rightarrow Ma'X) \]

\[ qY \rightarrow Ma'Z \sim Bnl \left( \frac{f'Ma'Z}{f'Ma'Z + f'Mb'X + f'Mb'Y + f'Mb'Z}, pY - (qY \rightarrow WX) - (qY \rightarrow WY) \right) - (qY \rightarrow WZ) - (qY \rightarrow Ma'X) - (qY \rightarrow Ma'Y) \]

\[ qY \rightarrow Mb'X \sim Bnl \left( \frac{f'Mb'X}{f'Mb'X + f'Mb'Y + f'Mb'Z}, pY - (qY \rightarrow WX) - (qY \rightarrow WY) - (qY \rightarrow WZ) \right) - (qY \rightarrow Ma'X) - (qY \rightarrow Ma'Y) - (qY \rightarrow Ma'Z) \]

\[ qY \rightarrow Mb'Y \sim Bnl \left( \frac{f'Mb'Y}{f'Mb'Y + f'Mb'Z}, pY - (qY \rightarrow WX) - (qY \rightarrow WY) - (qY \rightarrow WZ) \right) - (qY \rightarrow Ma'X) - (qY \rightarrow Ma'Y) - (qY \rightarrow Ma'Z) - (qY \rightarrow Mb'X) \]

\[ qY \sim pY' - (qY \rightarrow WX) - (qY \rightarrow WY) - (qY \rightarrow WZ) - (qY \rightarrow Ma'X) - (qY \rightarrow Ma'Y) - (qY \rightarrow Ma'Z) - (qY \rightarrow Mb'X) - (qY \rightarrow Mb'Z) \]

6-25

\[ qMa \rightarrow WX \sim Bnl (f'WX, \quad pMa) \]

\[ qMa \rightarrow WY \sim Bnl \left( \frac{f'WY}{f'WY + f'WZ + f'Ma'X + f'Ma'Y + f'Ma'Z + f'Mb'X + f'Mb'Y + f'Mb'Z}, pMa \right) - (qMa \rightarrow WX) \]
\[ qMa \rightarrow WZ \sim Bn! \left( f'WZ + f'MaX + f'MaY + f'MaZ + f'MbX + f'MbY + f'MbZ' \right) pMa \\
\quad - (qMa \rightarrow WX) - (qMa \rightarrow WY) \]

\[ qMa \rightarrow MaX \sim Bn! \left( f'MaX \right) \left( f'MaX + f'MaY + f'MaZ + f'MbX + f'MbY + f'MbZ' \right) pMa \\
\quad - (qMa \rightarrow WX) - (qMa \rightarrow WY) - (qMa \rightarrow WZ) \]

\[ qMa \rightarrow MaY \sim Bn! \left( f'MaY \right) \left( f'MaY + f'MaZ + f'MbX + f'MbY + f'MbZ' \right) pMa - (qMa \rightarrow WX) \\
\quad - (qMa \rightarrow WY) - (qMa \rightarrow WZ) - (qMa \rightarrow MaX) \]

\[ qMa \rightarrow MaZ \sim Bn! \left( f'MaZ \right) \left( f'MaZ + f'MbX + f'MbY + f'MbZ \right) pMa - (qMa \rightarrow WX) - (qMa \rightarrow WY) \\
\quad - (qMa \rightarrow WZ) - (qMa \rightarrow MaX) - (qMa \rightarrow MaY) \]

\[ qMa \rightarrow MbX \sim Bn! \left( f'MbX \right) \left( f'MbX + f'MbY + f'MbZ \right) pMa - (qMa \rightarrow WX) - (qMa \rightarrow WY) \\
\quad - (qMa \rightarrow WZ) - (qMa \rightarrow MaX) - (qMa \rightarrow MaY) - (qMa \rightarrow MaZ) \]

\[ qMa \rightarrow MbY \sim Bn! \left( f'MbY \right) \left( f'MbY + f'MbZ \right) pMa - (qMa \rightarrow WX) - (qMa \rightarrow WY) - (qMa \rightarrow WZ) \\
\quad - (qMa \rightarrow MaX) - (qMa \rightarrow MaY) - (qMa \rightarrow MaZ) - (qMa \rightarrow MbX) \]

\[ qMa \rightarrow MY \sim pMa - (qMa \rightarrow WX) - (qMa \rightarrow WY) - (qMa \rightarrow WZ) - (qMa \rightarrow MaX) \\
\quad - (qMa \rightarrow MaY) - (qMa \rightarrow MaZ) - (qMa \rightarrow MbX) \]

6-26

\[ qMb \rightarrow WX \sim Bn!(f'WX, \quad pMb) \]

\[ qMb \rightarrow WY \sim Bn! \left( f'WY \right) \left( f'WY + f'WX + f'MaX + f'MaY + f'MaZ + f'MbX + f'MbY + f'MbZ' \right) pMb \\
\quad - (qMb \rightarrow WX) \]
\[ \begin{align*}
qZ \rightarrow WZ \sim Bn & \left( \frac{f'WZ}{f'WZ + f'MaX + f'MaY + f'MaZ + f'MbX + f'MbY + f'MbZ}, pZ \right) \\
& - (qZ \rightarrow WX) - (qZ \rightarrow WY) \\
qZ \rightarrow MaX \sim Bn & \left( \frac{f'MaX}{f'MaX + f'MaY + f'MaZ + f'MbX + f'MbY + f'MbZ}, pZ - (qZ \rightarrow WX) \\
& - (qZ \rightarrow WY) - (qZ \rightarrow WZ) \\
qZ \rightarrow MaY \sim Bn & \left( \frac{f'MaY}{f'MaY + f'MaZ + f'MbX + f'MbY + f'MbZ}, pZ - (qZ \rightarrow WX) - (qZ \rightarrow WY) \\
& - (qZ \rightarrow WZ) - (qZ \rightarrow MaX) \\
qZ \rightarrow MaZ \sim Bn & \left( \frac{f'MaZ}{f'MaZ + f'MbX + f'MbY + f'MbZ}, pZ - (qZ \rightarrow WX) - (qZ \rightarrow WY) \\
& - (qZ \rightarrow WZ) - (qZ \rightarrow MaX) - (qZ \rightarrow MaY) \\
qZ \rightarrow MbX \sim Bn & \left( \frac{f'MbX}{f'MbX + f'MbY + f'MbZ}, pZ - (qZ \rightarrow WX) - (qZ \rightarrow WY) - (qZ \rightarrow WZ) \\
& - (qZ \rightarrow MaX) - (qZ \rightarrow MaY) - (qZ \rightarrow MaZ) \\
qZ \rightarrow MbY \sim Bn & \left( \frac{f'MbY}{f'MbY + f'MbZ}, pZ - (qZ \rightarrow WX) - (qZ \rightarrow WY) - (qZ \rightarrow WZ) \\
& - (qZ \rightarrow MaX) - (qZ \rightarrow MaY) - (qZ \rightarrow MaZ) - (qZ \rightarrow MbX) \\
qZ \rightarrow MY \sim pZ & - (qZ \rightarrow WX) - (qZ \rightarrow WY) - (qZ \rightarrow WZ) - (qZ \rightarrow MaX) - (qZ \rightarrow MaY) \\
& - (qZ \rightarrow MaZ) - (qZ \rightarrow MbX) - qMb \rightarrow MbZ \\
\end{align*} \]

\[ \begin{align*}
n'WX & = nWX - (qMa \rightarrow WX) - (qMb \rightarrow WX) - (qY \rightarrow WX) - (qZ \rightarrow WX) + (qW \rightarrow MaX) \\
& + (qX \rightarrow WY) + (qX \rightarrow WZ) + (qW \rightarrow MbX) \\
n'WY & = nWY - (qMa \rightarrow WY) - (qMb \rightarrow WY) - (qX \rightarrow WY) - (qZ \rightarrow WY) + (qW \rightarrow MaY) \\
& + (qY \rightarrow WX) + (qY \rightarrow WZ) + (qW \rightarrow MbY) \\
\end{align*} \]
Parameter selection

I ran simulations varying the ratio between the \( sa \) and \( sb \) selection coefficients by increasing the value of \( sb \) while keeping \( sa \) constant. This varies the interaction between the Ma mutant and the interfering Mb mutant. The simulation was stopped at 10,000 generations, when the WX type had become extinct in all simulations. The population size \( N \) was \( 10^6 \). Values for the rate of
recombination $\rho$ from 1 to 10000 were simulated. Values for the selection coefficient $s_a$ and $s_b$ were simulated from $10^{-3}$ to 0.25 (subject to the constraint that $s_a > s_b$).

6.4.2 Results

The results of a two locus model of clonal interference are shown in Figure 6.13. In this model a selective sweep is defined as occurring when the WX genotype is extinct in the population. At this point all the isolates in the population will contain either the Ma allele or the Mb allele. This study is concerned with the occurrence of polymorphism at the neutral locus and more specifically, the retention of neutral X alleles from the pre-sweep population, which can be incorporated into the final population through the formation of the MbX hybrid or the MaX hybrid. Once the WX genotype is extinct in the population no more X alleles can be rescued through recombination. Therefore the fate of the MbZ and MaY genotypes are not considered here.

![Figure 6.13](image.png)

Figure 6.13 Number of WX, MaY and MbZ in the population over time.

The results are shown for selection coefficients $s_a$ of 0.1 and $s_b$ of 0.005. In these simulations $\rho = 10$, $N=10^6$ and the results are presented for an average of 1000 simulations of the model.

As Figure 6.13 shows, the WX genotype is extinct in the population by 1500 generations and the MaY genotype makes up the majority of the population. In the previous section of this thesis I showed that an important factor in the conservation of wild-type neutral alleles from the pre-sweep population (X alleles) was the length of time the WX genotype remained in the population. I therefore examined whether the clonal interference increased the time to extinction of the WX
genotype. Figure 6.14(A) shows the result from two sets of simulations of the clonal interference model and a comparable set of simulations from the single advantageous allele model.

I then investigated how many alleles descended from the wild-type genotype in the pre-sweep population are retained in the population. As in the previous models an X allele is representative of any allele from a wild-type genotype in the pre-sweep population. In these simulations the X allele from a wild-type genotype can be retained in the post-sweep population by recombining with either advantageous mutant. The number of these bacteria which carries both an advantageous allele Ma and Mb and the X allele is shown in Figure 6.14(B). Figure 6.14(A) shows the number of wild-type bacteria (WX type) in a population of $10^6$ bacteria over 10,000 generations. Compared to a simple model with one advantageous allele (shown as a solid line), clonal interference has some effect on the length of time the wild-type genotype, WX, remains in the population. In the original model the WX type is completely extinct by 200 generations, in the clonal interference model with $sb = 0.01$ (shown as a dotted line) and $sb = 0.005$ (shown as a dashed line) extinction occurs at 400 and 1400 generations. In the original model WX falls below 10,000 bacteria at 188 generations, and in the clonal interference model at 420 generations with $sb = 0.01$. When $\rho = 100$, the clonal interference allows 700 more bacteria with the X allele to be retained in the population. In the context of the population as a whole ($10^6$) this is a very small proportional increase as shown in Figure 6.14 (B).
Figure 6.14 Dynamics of clonal interference over time. A) Total number of pure wild-type plotted and B) The number of new hybrids $MaX$ and $MbX$ plotted against time. The results are shown for selection coefficients $s_a$ of 0.1 and $s_b$ of 0.005 (dotted line) and $s_a$ of 0.1 and $s_b$ of 0.01 (dashed line), the solid line shows the result from the single mutant model when $s=0.1, \rho = 10$.

The simulations of the clonal interference extension to the model showed that advantageous mutations remain associated with the genetic background in which they originally occurred, with the exception of a small number of hybrids. The genetic diversity that had arisen before the occurrence of the two advantageous mutations is lost, therefore added complexity does not prevent a selective sweep reducing neutral diversity in the population.

### 6.5 Discussion and conclusion

In this section I will examine some key insights from the mathematical models presented in this chapter, and interpret these with reference to appropriate previous studies in published literature.

The occurrence of homologous recombination and the important role that it plays in shaping the genetic structure of bacterial populations is clear. However, the impact of recombination on different evolutionary processes is a source of study and debate (301). Using simple models I
investigated the impact of bacterial homologous recombination on a selective sweep which in the ecotype model provides a mechanism of speciation (259). The ecotype model groups bacteria into ecologically constrained monophyletic clusters. Selective sweeps allow the bacteria to adapt to their environment and restrict genetic diversity in the ecotype so that the ecotype remains a closely related cluster of bacteria (258, 259, 274). One of the objections to this model is that the ubiquity of homologous recombination threatens the formation of closely related genetic groups. Some investigators have maintained that selective sweeps are unable to occur in bacterial populations as homologous recombination would disrupt the process. They have stated that homologous recombination could re-assort the genotype faster than a selective sweep could be completed and so diversity would not be purged except at loci under selection (302, 278). I addressed this issue by introducing a model of selective sweeps incorporating bacterial homologous recombination which is suitable to analyse the impact of varying rates of localised recombination on a selective sweep. Using the stochastic two locus model I tracked the recombination events which retained neutral pre-sweep alleles. This showed that the speed of extinction of the wild-type population impacts on the amount of wild-type alleles retained, such that the latter was highest when the selection coefficient of the beneficial allele was small. This process was shown to be most important again in the simple model of clonal interference. In the simple two-locus model the proportion of wild-type alleles at the neutral locus was low when population recombination rates were <10 for selection coefficients for the beneficial allele at the selection locus of 0.01 or greater. Higher proportions could be rescued when recombination rates were much higher and this leads to a consideration of recombination rates in natural populations of bacteria.

i. Asymmetric bacterial recombination is not frequent enough in natural bacterial populations to prevent a selective sweep acting to remove genetic diversity over the whole genotype within an ecotype

Throughout this work I have used values of population recombination rates that were estimated from population genetic models which estimate the frequency of recombination relative to mutation (25, 15). In the first instance I considered the absolute population recombination rate rather than the ratio to mutation; these have been estimated as varying between 0.7, for Bacillus cereus, to 73.3 for H. pylori. These estimates are in agreement with empirical evidence (303) and certainly suggest that bacterial populations with population recombination rates above 50 are an exception (25, 15).

When examining the eight loci model I parameterised the population in terms of the ratio of recombination to mutation. I characterised a population with ρ/θ = 0.5 and 5 as being intermediately and highly recombining respectively. In terms of natural populations of bacteria, S. pneumonia has been estimated as having ρ/θ = 4 , N. meningitis ρ/θ = 9 and S. aureus ρ/θ = 0.5. These estimates
are taken from previous analysis of bacterial populations which have been demonstrated as being consistent with empirical evidence (25, 15). Another method of estimation of the ratio of recombination to mutation has been developed which has estimated much higher ratios of recombination to mutation from natural populations of bacteria. The results of this analysis reported more extreme ratios at both the recombining and clonal end of the spectrum. The authors computed a value called \( r/m \), which is the ratio of rates at which nucleotides become substituted as a result of recombination and mutation. If \( r/m = 10 \), then recombination introduces 10 times more nucleotide substitutions than point mutations during the evolution of the population this is consistent with a value of \( \rho/\theta = 2 \) if each recombination event introduces five substitutions on average (304). The authors reported a range of \( r/m \), from 63.3 to 0.02. Unfortunately the authors do not estimate absolute mutation or recombination rates and so it is difficult to interpret the results of the two locus model in light of these parameters. Interpreting the result of the eight locus model in light of these parameters, only four bacteria had an \( r/m \) above 25 (comparable to \( \rho/\theta = 5 \) (304). Thus the results of the model can be considered to hold for the majority of bacterial populations whose recombination rates have been estimated.

ii. The impact of a selective sweep is highly dependent on the selection coefficient

Although I have shown that homologous recombination cannot prevent the occurrence of a selective sweep in a bacterial population, the effect is highly dependent on the speed of the sweep. The speed of the sweep is in turn dependent on the relative fitness of the advantageous mutant allele. It remains unclear how common fast selective sweeps are in natural populations of bacteria and how useful the ecotype model is in describing the way in which most bacterial populations evolve. The range of fitness in laboratory populations of bacteria has been shown to be exponential in a well adapted population (284). Recent research has shown that this distribution is not an adequate description in populations under environmental stress when mutants of greater relative fitness are more common (305). This could indicate that a fast selective sweep will be more common in unstable environments.

iii. A monophyletic population is the predicted outcome of a selective sweep

These findings leads us to agree with one premise of the ecotype model, namely that a selective sweep will result in a monophyletic population. So within populations, or putative ecotypes, that are shaped by repeated selective sweeps, I would expect neutral variation and phenotypic characteristics to be aligned. A test of the ecotype model in various contexts is the extent to which such monophyletic associations are found in natural populations of bacteria. Although there has been substantial study of periodic selection in laboratory populations, there has been comparatively
little study of the relevance of this phenomenon in naturally occurring populations of bacteria. The study by Guttman and Dykuizen, 1994 examined the gapA gene (see the glossary in Appendix A for a brief explanation of gene names) in *E. coli* using nucleotide data. It found that the sequences of the gene were unusually alike. However other genes also sequenced in the study maintained diversity. The interpretation by the authors of the study is that a selective sweep occurred through the *E. coli* population at the gapA locus and recombination over the rest of the genome maintained diversity (306). This interpretation of the work has been questioned on the grounds that recombination is not sufficiently common in *E. coli* populations to maintain diversity during a selective sweep. Other researchers have interpreted the results of this study as showing distinct clusters of *E. coli* ecotypes in which separate periodic selection events of a universally advantageous allele have occurred (274). The advantageous mutant that is advantageous in all ecotypes had been passed from ecotype to ecotype by recombination. Of course, this interpretation can only be validated if corresponding ecotypes could be identified in the *E. coli* population, and this has not been done. In either case, the study only sequenced twelve isolates of *E. coli*, one of which was the laboratory strain K12. The small amount of isolates compared could leave the study vulnerable to sampling bias and a laboratory strain of *E. coli* is not representative of the evolution of a strain from a natural population of bacteria. A second study by Wagner and Riley identified low synonymous site variation at the lacY locus in *E. coli*. However, even if recombination could rescue diversity over the whole genome, it would be expected that recombination is extremely rare between neighbouring sequences. In the Wagner and Riley study the neighbouring phoA gene sequence was not found to share the same level of sequence diversity as the lacY locus (see the glossary in Appendix A for a brief explanation of gene cluster names) (307). If a selective sweep had acted on lacY I would expect low synonymous site variation at neighbouring genes such as phoA, as it is unlikely that close neighbours on the genome will be separated by recombination over the time it would take for a sweep to come to fixation.

In a review, Cohan describes a community phylogeny model constructed to test the ecotypes model (259). The community phylogeny model attempts to identify sequence clusters in natural populations which correspond to different ecotypes. This was applied to bacterial communities in the Negev Desert and identified a putative ecotype of *Bacillus licheniformis* specialised to the higher exposure to sunlight of the south facing slope of the canyon (259). The identification of clusters does not provide information about how those clusters developed and the selection processes involved and so does not necessarily provide evidence for selective sweeps in natural populations.
iv. A longitudinal sample of a bacterial population is needed to successfully examine the impact of a selective sweep

A more general issue raised by the results of this chapter is what metrics are appropriate for analysing the impact of selection on a bacterial population. Measures of allelic diversity such as Simpson’s $D$ and other measures of homozygosity are very useful in describing the overall reduction in genetic diversity in the population. What they cannot measure is the more subtle changes in the genealogy of the population, as the impact of the sweep on the amount of genetic difference in the population can be undone in a relatively short amount of time through mutation. A more lasting impact of that sweep is the change in the genealogy of the population. Using measures which quantify the proportion of wild-type genes remaining in the population and the probability of fixation of a wild-type gene, the ancestry of a post-sweep population can be traced to the first bacteria carrying the advantageous mutant. The impact of homologous recombination on the population is most keenly felt on the genealogy of the population (at the end of the sweep most bacteria are descended from the first bacteria carrying the advantageous mutant) rather than the genetic diversity which may be relatively quickly restored through mutation. Thus any study of selective sweeps in bacterial populations should be carried out using metrics which quantify both difference and descent as used in this study. These results also highlight the importance of using good longitudinal data to properly show selective sweeps in natural bacterial populations, rather than sampling a snapshot in time.
7 Impact of pneumococcal carriage vaccination on the population structure of Streptococcus pneumoniae

7.1 Introduction

The previous two chapters examined a simple selective sweep where a bacterium developed an advantageous mutation and consequently increased in number in a population. As I reviewed in Chapter Five this is the form of bacterial selection that has been most commonly studied using population genetic theory and in the laboratory. To use population genetic theory to understand the impact of human interventions in bacterial populations it is necessary to examine more complex selective scenarios. In 2000 the PCV7 vaccine was introduced in the United States, which placed a strong selective pressure on serotypes which the vaccine targeted, as I will review below. This chapter presents analysis of the impact of the PCV7 vaccination programme on genetic diversity (characterised using MLST) in the pneumococcal population.

i. Vaccination with PCV7 has resulted in serotype replacement in carriage

The introduction of the PCV7 vaccine has resulted in a reduction of IPD caused by the seven serotypes targeted by the vaccine (308). However, in the years following the introduction of the vaccine there has been an increase in IPD caused by serotypes not targeted by the vaccine, most notably serotype 19A (56). There has been concern that over time this serotype replacement will continue and erode the benefits of vaccination. In response to this Hanage et al. (309) examined a longitudinal carriage sample to examine whether the increase in serotypes not targeted by the vaccine could be expected to continue, or whether the frequency of serotypes has reached a new equilibrium and so future changes in serotype frequency will be small (309). Using population genetic techniques, the authors examined whether the impact of selection pressure from vaccination could be detected and whether the pneumococcal population returned to a dynamic equilibrium during the period studied. They compared the serotype diversity in samples collected from a longitudinally sampled vaccinated community with three communities sampled prior to the introduction of vaccination. The authors showed that there was a statistically significant increase in serotype diversity in the vaccinated communities compared to un-vaccinated communities indicating the selective pressure of vaccination. This is illustrated in Figure 7.1 which shows the relative proportions of serotype plotted by rank frequency. The sample collected in 2007 showed no significant difference in serotype diversity from the pre-vaccination population (309).
Figure 7.1 Rank frequency distributions of serotypes in the samples collected in Massachusetts during 1998-9 (A), 2001 (B), 2004 (C) and 2007 (D).

The median values obtained from 1000 samples of the same size drawn from the combined pre-vaccine distribution are shown as a solid line with dashed lines either side representing the 5th and 95th percentiles. Larger numbers of serotypes in the later samples reflect larger sample size. Figure reproduced from (309).

In the study conducted by Hanage et al., the authors considered that selective pressure of the vaccine will only impact on the specially targeted serotypes included in the vaccine. Before the introduction of the PCV7 vaccination, vaccination was thought to be protective against colonisation
by serotypes 6A, 9A, 9L, 9N, 19A, 23A and 23B, due to cross-reactive protection (48). These were referred to as PCV7 cross reactive serotypes. However, the response to vaccination in the prevalence of these serotypes has not been uniform. 6A, 9A and 9N have reduced in prevalence in carriage studies. Conversely 19A, 23A and 23B have increased in prevalence in carriage studies (5). Thus, a serotype’s putative cross reactivity in vitro has not been shown to be a useful predictor of their behaviour in PCV7 vaccinated communities. This is in contrast to those serotypes included in the PCV7 vaccination which have shown consistent reduction in incidence in response to vaccination (6).

ii. Vaccination pressure may result in sequence type replacement, though the extent may depend on the impact of capsular switching

As discussed in Chapter 1, the variation between strains in a pneumococcal population can be indexed by identifying antigenic (serotype) variation or genetic variation that is not under known selective pressure. As vaccination targets certain serotypes, the indexing of antigenic variation is an essential part of understanding the impact of the vaccination campaign. However, the identification of capsular switching and the transfer of antibiotic resistant genes (described in Chapter 1) shows that important clinical characteristics, such as antibiotic resistance, may be associated with particular genotypes rather than a serotype and highlight the importance of identifying and understanding changes in genetic variation within a serotype.

As the current pneumococcal vaccines target the serotype, the wider genome is not under selective pressure from the vaccine. The ST genes do not include the capsular polysaccharide genes and therefore are a measure of genetic diversity in parts of the genome not under direct selective pressure from vaccination. However, a simple model of selection would predict that STs associated with the vaccine serotypes would decline in frequency in the population in response the selective pressure of the vaccine. As serotypes not targeted by the vaccine increase in frequency in the population, STs associated with these serotypes would also increase in frequency in the population. It is therefore of interest whether strain structure changes in response to vaccination as expected, or whether certain STs are more ‘successful’ than others and expand in the population more than expected.

Although there is a strong association between specific serotypes and associated STs, capsular switching has been documented in pneumococcal populations (51), as described in Chapter 1. Therefore STs that were associated with vaccine serotypes may remain in a pneumococcal population under selective pressure from vaccination, though now associated with a non-vaccine
serotype. Thus, although serotype replacement may occur in the pneumococcal population, the extent of associated ST replacement is dependent on the impact of capsular switching.

7.1.1 Aims

In an extension of the analysis carried out by Hanage et al. (309) I will examine the impact of PCV7 vaccination on genetic variation and population structure of strains in pneumococcal carriage. To this end I will analyse samples of carried pneumococci from Massachusetts, the UK and Finland to:

- Examine whether vaccination results in ST replacement,
- Analyse the impact of vaccination on ST frequency and diversity,
- Identify the impact of capsular switching if it is present.
7.2 Methods

7.2.1 Data-sets

i. Un-vaccinated communities

I used three data-sets collected from un-vaccinated communities. As an indicator of baseline ST distribution in Massachusetts I analysed a sample of 71 carried isolates from children aged between three months and six years of age. The samples were collected between 1998 and 1999 and all children in the sample had attended a family practice suffering from acute otitis media (AOM) (46). These isolates were assumed to be broadly representative of the distribution STs in nasopharyngeal carriage in the community during this period.

As there may be small differences in the ability of carried serotypes to cause AOM (310), I also analysed two samples from un-vaccinated communities outside Massachusetts. I analysed a sample of carried isolates from healthy children less than two years of age in Oxford, UK (311) and Tampere, Finland (310).

Isolates in the UK data-set had been obtained from a longitudinal pneumococcal carriage study performed by the Oxford vaccine group (University of Oxford) during 1999-2001. Children were enrolled in this study at birth and lived in the city of Oxford and surrounding area. Samples of nasopharangeal flora were obtained from 100 children sampled repeatedly over two years. 264 isolates obtained and characterised by MLST were used in this data-set (311).

The Finnish data-set was collected from a survey of 329 children from Tampere, Finland. The children were monitored from 2 to 24 months of age as part of a study into acute otitis media (FinOM) (312). S. pneumoniae were isolated from nasopharyngeal samples from healthy children in the FinOM cohort study between 1994 and 1997 (310).

All three data-sets have been modified by the authors of the published studies so that each isolate was from one episode of carriage per child. Where more than one isolate with the same ST and serotype was collected from the same child this was excluded from the data-set. The size of the data-sets used in the analysis are MA 1998-8 N=71, Finland N=216 and UK N=228.
Table 7.1 Original data-set composition of un-vaccinated communities

<table>
<thead>
<tr>
<th>Location of study</th>
<th>UK (311)</th>
<th>Finland (310)</th>
<th>MA 1998-9 (46)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of isolates</td>
<td>264</td>
<td>217</td>
<td>90</td>
</tr>
<tr>
<td>Age of children</td>
<td>0-24 months</td>
<td>2-24 months</td>
<td>Under 7 years</td>
</tr>
<tr>
<td>Type of study</td>
<td>Birth cohort</td>
<td>Birth cohort</td>
<td>Antibiotic resistance survey</td>
</tr>
</tbody>
</table>

**ii. Longitudinal sampling in vaccinated communities**

Samples were collected from children under seven years of age in 2001, 2004, 2007 and 2009 from children attending family practices in eight Massachusetts communities. In 2001 and 2004, 16 communities were sampled; the eight additional communities sampled during these years were excluded from this analysis for consistency. Details of the collection of samples from these communities in 2001, 2004 and 2007 is described in detail in (48,47,313,50). Samples were collected in 2009 following a method consistent with previous years described in references (48,47,313,50).

I will briefly summarise the process of data collection. The communities used in the studies were selected on the basis of geographic separation, evidence that few children crossed community boundaries to seek paediatric care, and diversity of size and demographic characteristics using available US census data (48). In all data-sets isolates were collected from young children under seven years of age. This age group was chosen to include the period of highest risk of pneumococcal carriage and to be large enough to collect the required sample size. Children were eligible for inclusion if they were younger than 7 years, resided in a study community, and presented for either a routine well-child or “sick visit” at a participating practice. Parental consent and nasopharyngeal swabs were obtained by either trained study personnel or trained office practice nurses. PCV7 vaccination history was recorded from the medical record (30).

Specimens were collected from children using a calgiswab which was inserted into the back of the nose of the child. The sample was placed into transport medium and transported for analysis at the Children’s Hospital in Boston (30). *S. pneumoniae* colonies were identified from nasopharyngeal samples through growth on specific media. Serotype was determined for a single colony from a pure culture of each isolate by the quellung reaction, using serotype-specific pneumococcal antisera (30).

The assignment of STs to isolates in the data-sets used in this analysis has previously been described in detail (313). In summary, the ST of each isolate was determined by MLST as previously described.
For each of the seven housekeeping genes (loci); *aroE, gdh, gki, recP, spi, xpt and ddl*, nucleotide sequences were obtained on both DNA strands. Alleles and ST assignments were made using the MLST website (www.spneumoniae.mlst.net). A summary of the data-sets is presented in Table 7.2.

7.2.2 Analysis

i. Estimation of ST diversity

The diversity of each data-set was assessed using Simpson’s index of diversity $D$ (292). This has been used previously to describe changes in serotype diversity in the samples described in this chapter. Here I use it to examine ST diversity and so Simpson’s $D$ is the probability that two isolates chosen at random will have a different ST. The standard calculation of Simpson’s $D$ includes self pairs if calculated in a finite population, which in a small population may reduces the usefulness of the measure. Simpson’s $D$ corrected for a finite population size is defined as:

$$D = \frac{1 - \sum x_i^2}{1 - \frac{1}{N}}$$

The fraction with ST $i$ is defined as:

$$x_i = \frac{n_i}{N}$$

In these calculations $m$ is the total number of STs and $N$ is the total sample size. The variance and 95% confidence interval were calculated as previously (for an infinite population due to computational necessity) (315). The values of $D$ will be distributed around the true diversity with the variance $\sigma^2$ defined as:

$$\sigma^2 = \frac{\frac{9}{N} \left[ \sum x_i^3 - \left( \sum x_i^2 \right)^2 \right]}{N}$$

An estimate of the standard deviation of the true diversity of a population is given by $\sqrt{\sigma^2}$ and the approximate 95% confidence intervals estimated as:

$$CI = \left[ D - 2\sqrt{\sigma^2}, D + 1.96\sqrt{\sigma^2} \right]$$ (315).
To estimate whether the values of $D$ were significantly different between the data-sets I calculated a $P$ value using a two tailed Welch’s t-test.

To further explore the impact of vaccination I calculated the rank frequency distribution of STs for each sample. The frequency of each ST is described by equation 7-2. Having calculated the distribution of frequency of STs denoted by $x_1, x_2, ..., x_m$ I ranked the frequencies from largest (rank 1) to smallest (rank $m$), where $m$ is the total number of STs in the sample.

To compare a range of rank frequency distributions to the vaccine era samples, I simulated 1000 Monte Carlo replicates drawn from the empirically observed rank-frequency distribution of the non-vaccine sample MA 1998-1999. This analysis was carried out using the R statistical package.

ii. Analysing changes in STs using eBurst

To examine the longitudinal trends in STs and their association with serotypes in a community I analysed data-sets collected in 2001, 2004, 2007 and 2009 using eBurstV3 software (316). This software analyses the allelic profiles in a data-set and groups isolates into eBurst groups. eBURST groups are considered to approximately define clonal complexes, STs that share sufficient similarity that they are considered to descend from a recent common ancestor. In this analysis an ST was assigned to a clonal complex when identical at $\geq 6$ loci to at least one other ST in the group. Details of the number, size and makeup of clonal complexes for each data-set was generated. A population snapshot (a diagram displaying all clonal complexes and singletons in the sample) was also generated for each data-set.

I then divided the sample into vaccine and non-vaccine serotypes. The allelic profiles of the vaccine associated sub-set were compared to the allelic profiles of the non-vaccine associated sub-set using the eBurst comparison tool. I generated a population snapshot which compared the STs found associated with vaccine serotypes with those STs found associated with non-vaccine serotypes.

7.3 Results

7.3.1 ST diversity

i. ST diversity in un-vaccinated communities

To analyse ST diversity in carriage prior to vaccination, I examined independent samples from similar age groups in three un-vaccinated communities, presented in Table 7.1. Displayed in Table 7.2 estimates of Simpson’s $D$ calculated from the three data-sets. The sample collected in Massachusetts prior to vaccination has a smaller measure of Simpson’s $D$ (lower diversity) than the samples.
collected in Finland and UK. The differences between Simpson’s $D$ are quite substantial between values close to 1 (the absolute maximum), however, there is no statistically significant difference between these values of Simpson’s $D$.

Figure 7.2(A-C) shows the rank frequency of the STs plotted for each sample collected in un-vaccinated communities. For the Tampere (B) and UK (C) samples, the MA 1998-1999 distribution is shown for comparison and the 95% intervals of 1000 samples drawn from the empirically observed MA 1998-1999 rank frequency distribution are also shown.

The Massachusetts data-set sampled prior to vaccination (MA 1998-9) is the least diverse of the three populations. In this data-set the most common and second most common STs makes up 18% of the population. In the UK and Finland data-set the two most common STs make up 12% and 11% of the population respectively.

Table 7.2 Sample characteristics of vaccinated and unvaccinated communities

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample Size</th>
<th>Number of STs</th>
<th>$D$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA 1998-9</td>
<td>71</td>
<td>43</td>
<td>$0.97 (0.96, 0.98)$</td>
</tr>
<tr>
<td>Finland</td>
<td>216</td>
<td>91</td>
<td>$0.98 (0.98, 0.99)$</td>
</tr>
<tr>
<td>UK</td>
<td>228</td>
<td>100</td>
<td>$0.98 (0.97, 0.98)$</td>
</tr>
<tr>
<td>MA 2001</td>
<td>126</td>
<td>61</td>
<td>$0.97 (0.96-0.99)$</td>
</tr>
<tr>
<td>MA 2004</td>
<td>222</td>
<td>85</td>
<td>$0.96 (0.95-0.98)$</td>
</tr>
<tr>
<td>MA 2007</td>
<td>291</td>
<td>86</td>
<td>$0.96 (0.95-0.97)$</td>
</tr>
<tr>
<td>MA 2009</td>
<td>293</td>
<td>94</td>
<td>$0.97 (0.97-0.98)$</td>
</tr>
</tbody>
</table>
Figure 7.2 Frequencies of individual sequence types (STs) in data-sets sampled from unvaccinated communities arranged by rank. (A) MA 1998-1999, (B) Finland and (C) UK. In graphs (B) and (C) the rank frequency distribution in the baseline unvaccinated community MA1998-9 is plotted as a blue dashed line. The 2.5th and 97.5th percentile values for each point obtained from 1000 samples of the sample size drawn from the baseline unvaccinated community (MA 1998-1999) are shown as a blue shaded section in graphs (B) and (C).
ii. **Comparison of ST diversity in un-vaccinated and vaccinated communities**

The values for Simpson’s $D$ were calculated for each sample. The lowest value of $D$ was 0.96 and estimated from MA 2004, 0.96 (95% CI 0.95-0.98) and MA 2007, 0.96 (95% CI 0.95-0.97). The value of $D$ in the pre-vaccine population in Massachusetts (MA 1998-9) was 0.97 (95% CI 0.95, 0.99) and by 2009 the estimate of $D$ had returned to this level. However, none of these changes in value were statistically significant at the 5% level.

Figure 7.3 (A-D) shows the rank frequency distribution is for the post-vaccination samples and the MA 1998-9 sample, which is plotted as a blue broken line for the purposes of comparison in each graph. The MA 1998-1999 sample was chosen as the baseline population. For each sample the 5% and 95% percentile values of 1000 samples drawn from the empirically observed MA 1998-1999 rank frequency distribution are also shown.

In the samples collected immediately after the implementation of vaccination (MA 2001) ST diversity decreased in the population compared to the sample collected prior to vaccination. The peak homozygosity of the population was reached in 2004 and 2007 when the most common ST made up 14% of the population. In the samples collected in 2009 the ST diversity is increased and by 2009 the most common ST made up 8% of the population. This sample was not significantly different to the baseline estimate taken from the MA 1998-9 population when the most common ST made up 9% of the population.
Figure 7.3 Rank-frequency distributions of individual sequence types (STs) in data-sets sampled from vaccinated communities MA 2001 (A), MA 2004, (B), MA 2007 (C) and MA 2009 (D). The rank frequency distribution in the baseline un-vaccinated community MA1998-9 is plotted as a blue dashed line in each. The 2.5th and 97.5th percentile values for each point obtained from 1000 samples of the sample size drawn from the baseline unvaccinated community (MA 1998-1999) are shown as a blue shaded section in each graph.
7.3.2 Changes in STs after the introduction of vaccination

i. eBurst analysis

In order to examine the effect of vaccination on the diversity of STs in the population, I first established which STs were associated with serotypes included in the vaccine. At the time of writing the full information on the serotype of each isolate in the data-set collected in Massachusetts in 1998 was not available. Therefore for this analysis I used 2001 as my baseline population.

For 2001, 2004, 2007 and 2009 I compared isolates with a PCV7 vaccine serotype (4, 9V, 14, 19F, 23F, 18C, 6B) and isolates with other serotypes using an eBURST comparative population snapshot. In this eBurst population snapshot each ST is represented by a yellow circle. The size of the circle is the relative frequency of the ST in the sample. The larger the circle the greater the number of STs. When an ST differs at one locus from another ST it is referred to as a single locus variant (SLV) and a black line is drawn between the two STs. In a clonal complex the probable ancestor of the clonal complex is assumed to be the isolate with the greatest number of SLVs and is coloured blue. In these comparative population snapshots, an ST number with a number labelled in green is only associated with vaccine type (VT) serotypes, those in black with non-vaccine type (NVT) serotypes, and those in pink were found with both VT and NVT serotypes.

Figure 7.4 shows the population snapshot for the sample collected in 2001. In this sample only two STs are associated with both a VT and NVT serotype (ST 439 and ST 199). All other STs are associated with either a VT or NVT serotype.

Figure 7.5 shows analysis of the 2004 population. There has been an increase in STs that were not associated with NVT serotypes in the 2001 sample. Of particular note are STs 62 and 558 which have increased in frequency in the population. In Figure 7.4 and Figure 7.5 these STs are circled in red. STs 199 and 439, which had previously been associated with both serotype groups, are now associated solely with NVT serotypes. ST 320, previously associated with only VT serotypes, is now only associated with NVT serotypes. The population snapshot from 2007 (Figure 7.6) shows that STs associated with NVT serotypes expand in the population but, other than STs 199 and 439, do not form clonal complexes. By 2009 (Figure 7.7) STs 62, 439 and 38 have formed clonal complexes.
Figure 7.4 Population snapshot of sequence types in 2001.
The green labelled sequence types are associated with PCV7 serotypes. The black labelled sequence types are associated with non vaccine serotypes. STs 199 and 439 are found associated with both serotype groups and coloured in pink. STs 558 and 62 referred to in the text are circled in red.
Figure 7.5 Population snapshot of sequence types in 2004. The green labelled sequence types are associated with PCV7 serotypes. The black labelled sequence types are associated with non vaccine serotypes. STs 448, 667 and 690 are found associated with both serotype groups and coloured in pink. STs 558 and 62 referred to in the text are circled in red.
Figure 7.6 Population snapshot of sequence types in 2007.
The green labelled sequence types are associated with PCV7 serotypes. The black labelled sequence types are associated with non vaccine serotypes. STs 138, 23 and 42 are found associated with both serotype groups and coloured in pink. STs 558 and 62 referred to in the text are circled in red.
Figure 7.7 Population snapshot of sequence types in 2009.
The green labelled sequence types are associated with PCV7 serotypes. The black labelled sequence types are associated with non vaccine serotypes. STs 1390 and 36 are found associated with both serotype groups and coloured in pink. STs 558 and 62 referred to in the text are circled in red.
**ii. Impact of serotype replacement on ST frequency**

Analysis of the longitudinal samples of vaccinated communities using eBurst demonstrated that the ST composition changed over time. There was a notable increase in those STs associated with NVT serotypes and a decrease in STs associated with VT serotypes. However some STs which had been associated with VT serotypes remained in the population. Table 7.3 examines the fate of STs associated with VT serotypes in more detail. In 2001 30% of STs in the population were associated only with a VT serotype. These STs are labelled with a green ST number in Figure 7.4. Only 5% of that 30% is still present in 2007. Three STs were associated only with VT serotypes in 2001 and survived into 2007: STs 138, 320 and 36 (shown in Table 7.4). The rise in these STs to make up 14% of the STs in the 2009 population is a result of the expansion of ST 320. By this time ST 320 is associated with a NVT serotype 19A.

**Table 7.3 Frequency of STs associated with VT and NVT serotypes in 2001 and change in frequency in later years**

<table>
<thead>
<tr>
<th>Year</th>
<th>VT associated in 2001&lt;sup&gt;a,b&lt;/sup&gt;</th>
<th>NVT associated in 2001</th>
<th>Shared associated in 2001&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>0.30</td>
<td>0.56</td>
<td>0.14</td>
</tr>
<tr>
<td>2004</td>
<td>0.07</td>
<td>0.77</td>
<td>0.16</td>
</tr>
<tr>
<td>2007</td>
<td>0.05</td>
<td>0.79</td>
<td>0.15</td>
</tr>
<tr>
<td>2009</td>
<td>0.14</td>
<td>0.88</td>
<td>0.12</td>
</tr>
</tbody>
</table>

<sup>a</sup> STs associated only with 4, 9V, 14, 19F, 23F, 18C, 6B serotypes in 2001  
<sup>b</sup> In 2004, 2007 and 2009 this is the proportion of STs which had been associated with a VT in 2001 (may now be associated with a NVT)  
<sup>c</sup> Sequence types associated with vaccine serotypes and one other non-vaccine serotype in 2001

**Table 7.4 Capsular Switching**

<table>
<thead>
<tr>
<th>ST</th>
<th>2001</th>
<th>2004</th>
<th>2007</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>138</td>
<td>6B</td>
<td>6B</td>
<td>6C &amp; 6B</td>
<td>6C &amp; 6A</td>
</tr>
<tr>
<td>320</td>
<td>19F</td>
<td>19A</td>
<td>19A</td>
<td>19A</td>
</tr>
<tr>
<td>36</td>
<td>23F</td>
<td>23F</td>
<td>19A &amp; 23B</td>
<td>23B &amp; 14</td>
</tr>
</tbody>
</table>
7.4 Discussion

i. Impact of selection on genetic diversity

I found that I could detect the impact of selection, due to vaccination, on the population structure of carried pneumococci by comparing three longitudinal MLST samples collected from vaccinated communities with samples collected in communities prior to vaccination. There were notable changes in the distribution of STs in vaccinated communities when analysed using the rank frequency distribution. However unlike analysis of serotype diversity published previously, changes in the estimated value of Simpson’s $D$ were not statistically significant, results are presented in Table 7.2.

The rank frequency distribution of STs in the longitudinal samples from 2001, 2004, 2007 and 2009 document changes in the population structure in response to conjugate vaccination. Between 2001 and 2004 the population is less diverse, while this trend is reversed between 2004, 2007 and 2009. This is comparable with previous research into the effect of vaccination on serotype diversity (317) and consistent with vaccination rates in the population. Vaccine coverage in the Massachusetts area has been documented as low during 2001 and 2002. In the 2001 data-set 36% of children aged 19-35 months had received three doses of vaccine (10). When the sample was collected from this age group in 2004 90% coverage by vaccination had been achieved (48). It is a reasonable assumption that the decrease in diversity seen in 2004 reflects the rapid expansion of a few STs associated with non-vaccine serotypes.

ii. By 2009 genetic diversity has returned to pre-vaccination levels

By 2009 the distribution of STs in the population has regained the shape characteristic of an un-vaccinated population. It is reasonable to assume that between 2004 and 2009 previously rare STs become frequent in the population and novel SLVs of these STs are generated through mutation and recombination, increasing ST diversity. This assumption will be examined in more detail in the following sections. In a recently published study Hanage et al. suggested that the changes in serotype diversity seen in 2004 reflect the point at which the pneumococcal population was in the process of responding to selective pressure imposed by the pneumococcal vaccine (3). In their analysis they showed that by 2007 the serotype distribution had returned to the shape and diversity seen in un-vaccinated communities, though the serotype composition was now very different (3). The authors predicted that no further expansion of non-vaccine serotypes would be expected. The
results from the 2009 sample certainly suggest that the authors were correct in their prediction and the pneumococcal population in Massachusetts has reached a post-vaccination dynamic equilibrium.

iii. **Changes in ST frequency in response to vaccination**

In the sample collected in 2001, the majority of STs are associated solely with either a VT or NVT serotype, though each serotype can have more than one ST associated with it. Through 2004, 2007 and 2009 the frequency of STs associated with VT serotypes declines as the frequency of VT serotypes decline, which is consistent with studies that have documented the impact of serotype replacement in this community. In 2004, 2007 and 2009 STs which had been seen at low frequency in the sample collected in 2001 and associated with NVT serotypes increase in frequency in the population, to the exclusion of STs associated with NVTs. By 2007 only 5% of STs associated with VT serotypes in 2001 were still present in the population, in contrast 80% of the STs associated with NVT serotypes in 2001 were still present in the population. Thus, serotype replacement was accompanied by a noticeable ST replacement, though some STs common to both VT and NVT serotypes (e.g. ST 199 and ST 439) persist at high frequency in all samples.

iv. **Impact of capsular switching**

A small percentage of STs that were associated only with VT serotypes in 2001 were still present in the population in 2007. In 2007 these STs were no longer associated with VT serotypes. STs 138, 320 and 36 are associated with a NVT serotype in 2007. In 2009 ST 36 is also found associated with VT serotype 14, which it had not been associated with previously. In the case of each of these STs the apparent novel combination could have arisen through recombination (capsular switching) in the recent past or this combination of ST and serotypes could have been present in pneumococcal populations previously. As there is no record of the ST serotype combination in the global MLST database (http://spneumoniae.mlst.net) prior to 2000, it is reasonable to assume that these combinations of ST and serotype have arisen post-vaccination, as has been documented previously (18). Therefore, due to the role of capsular switching, ST replacement was not as comprehensive as serotype replacement in the pneumococcal population.

ST 320 is of particular public health interest as this ST is associated with multi-drug resistance. ST 320 was associated with serotype 23F (a VT) in 2001 and is associated with 19A (a NVT) in 2004, 2007 and 2009. Indeed, between 2007 and 2009 it increases in frequency in the population, while the prevalence of 19A remains relatively constant in those two years. This suggests that although the frequency of serotypes may now be at a stable equilibrium, changes in the constituent STs of each serotype may still occur.
v. Limitations of analysis

There are a number of limitations in this analysis. The sample size is variable across data-sets and the sample size of the 2001 and 2004 data-set is not very large. This is due to the exclusion of data collected in eight additional communities that were not sampled in 2007. In addition, the sample from an un-vaccinated community in Massachusetts was collected from nasopharyngeal carriage in children with AOM. The serotypes and thus sequence types which cause AOM may not be representative of those found in carriage. The differences between serotypes causing AOM and those found in carriage are expected to be minor (as indicated by analysis presented by Hanage et al. (6)). This data-set is only an approximate representation of the pneumococcal population in Massachusetts prior to vaccination and so direct comparison of ST identity with samples from 2001, 2004, 2007 and 2009 is not appropriate.

In the first instance change in ST diversity was assessed using estimates of Simpson’s $D$. This measure did not find any significant change in the diversity in the populations over time. The lack of significant change could be due to the small size of the samples, or the measure may not be suitable to assess ST diversity from these data-sets.

The simple selective sweep model with levels of recombination seen previously in pneumococcal populations predicts that only a small proportion of STs in the post-vaccination population would have originally been exclusively associated with vaccine serotypes. In general we would expect that serotype replacement should result in associated ST replacement. Table 7.3 shows that this expectation is broadly validated. In 2001 30% of STs were exclusively associated with vaccine serotypes. In 2007 this had dropped to 5%. Shown in Table 7.4, the STs that survived did so as a result of serotype switching. STs 138, 320 and 36 recombined with a non-vaccine serotype which they are now associated with, allowing them to persist in the population. This phenomenon is known as capsular switching and has been documented previously (317).

The STs which switch capsules are more likely to come from serotypes that persist for longer at intermediate levels in the population. This expectation is explained in Chapter 5 of the thesis where Figure 5.2 illustrates that recombination events are more likely to occur when the wild-type and mutant type genomes are at intermediate levels. This is validated in the pneumococcal population. The STs which recombined with a non-vaccine serotype were previously associated with the most common vaccine serotypes; 19F, 6B and 23F which were also serotypes that persisted in the population (46, 47).
The selective sweep model and this analysis assumed that STs are selectively neutral: no one ST is more likely to persist in the population than others and the STs that persist through recombination do so through chance only. Other authors have postulated that some STs are selectively advantageous due to multiple antibiotic resistant genes or other unknown characteristics associated with the ST (41,51), if STs were not selectively neutral then a larger proportion would survive than predicted by the simple sweep model. This analysis found that 5% of STs persisted in the population that were previously exclusively associated with PCV7 serotypes. This is within the maximum prediction of the simple selective model which assumes that all STs are neutral. However given that capsular switching is likely to be less common than rates of recombination simulated in the simple selective sweep model and that the subsequent increase in STs (particularly ST 138) previously associated with PCV7 serotypes, the results of this model do not fully conform to the neutral expectations.

7.5 Conclusion

I have used a number of simple population genetic tests and frequency distributions to detect the impact of vaccination on a carried population of pneumococci. I demonstrated that the ST diversity in carried pneumococci decreased after vaccination and by 2004 changes in population structure are evident. However, by 2009 the distribution of ST frequencies is very similar to that seen in an un-vaccinated population. I have also demonstrated that ST replacement occurred over this time, although ST replacement was not complete due to the impact of capsular switching.
8 Discussion

Discussions of detailed findings are included in the relevant chapters. This chapter provides an overall summary of key findings from each chapter, describes the implications of these findings for pneumococcal treatment and prevention and identifies research which follows from this thesis.

8.1 Summary of findings

The first chapter introduced *S. pneumoniae* as a pathogen and a commensal of the upper respiratory tract, described its transmission and life cycle and outlined where the greatest burden of morbidity and mortality lay. It also outlined that due to enhanced surveillance of pneumococcal invasive disease and studies of pneumococcal carriage there was now an unprecedented wealth of data describing the serotype distribution, antibiotic resistance prevalence and population structure of *S. pneumoniae*. It outlined that there remained challenges to understand this population structure and showed that previous research into multi-strain systems did not adequately address the research questions outlined in the aims of this thesis.

In Chapter 2 I performed a meta-analysis to examine the impact of antibiotic non-susceptibility on the risk of mortality. In contrast to previous reviews on this subject, I examined the impact of antibiotic non-susceptibility in pneumococcal pneumonia, bacteraemia and meningitis. I showed that PNSP pneumococcal pneumonia or bacteraemia infections did not have an increased mortality associated with penicillin treatment. However, a meningitis infection with a PNSP pneumococci increased the risk of mortality two fold, when these patients were treated with penicillin. I concluded that penicillin could still be used to successfully treat many non-meningeal pneumococcal infections. These results concur with the recent revision of appropriate Penicillin MICs for the treatment of pneumonia and bacteraemia (65) and medical opinion that penicillin levels in the blood and lungs are sufficient to kill PNSP pneumococci but that the blood-brain barrier results in much lower levels in the CSF and penicillin cannot reliably be used to treat non-susceptible pneumococcal CSF infections (111,113).

Chapter 2 demonstrated the impact of the loss of effective treatment due to the increase in antibiotic resistance. In Chapter 3 I examined why some serotypes of the pneumococcus are more likely to carry resistance than others. Using a mathematical model I generated the hypothesis that serotypes with a long duration of carriage are more likely to have a high prevalence of antibiotic resistance than those with a short duration of carriage. In Chapter 4 I tested this hypothesis by
analysing the prevalence of penicillin resistant isolates of *S. pneumoniae* collected as part of invasive disease surveillance. Using maximum likelihood estimation, I showed that in children less than two years of age, penicillin resistance only occurred in serotypes which had duration of carriage of nineteen days or more.

In Chapters 3 and 4 I considered the circumstances under which antibiotic resistance carries a selective advantage in a pneumococcal serotype. In Chapters 5 and 6 I considered the effect that the spread of an advantageous gene has on the genetic diversity in a generic *in silico* bacterial population. Specifically, I considered whether asymmetric recombination typical of recombination can prevent the loss of diversity over the whole genome.

In Chapter 5 I reviewed the literature in this area, and compared the results of models of selection incorporating asymmetric and symmetric recombination. From the results of this analysis, I showed that a mathematical model incorporating asymmetric recombination should be used to examine the impact of selection on a bacterial population. In Chapter 6, I developed a stochastic two locus model and a multi loci model of bacterial evolution which examined whether recombination can rescue genetic diversity associated with the ‘wild-type’ phenotype present in the population before the selective sweep. I explored a range of parameters and when I parameterised the model using recombination rates estimated from natural populations of bacteria, the loss of genetic diversity was a consequence of a selective sweep. Diversity at loci that are not under selection is lost due to the hitch-hiking effect of a selective sweep, when neutral loci associated with the locus under selection sweep with the advantageous allele to fixation in the population.

In Chapter 7, I examined the impact of selective pressure from the conjugate vaccines on genetic diversity in a pneumococcal population. I analyse a longitudinally sampled MLST dataset of carried isolates. These isolates were collected in 2001, 2004, 2007 and 2009 from healthy children, and using the rank frequency distribution successfully detected the impact of selection by the conjugate vaccination. I show that in the period immediately following the introduction of vaccination sequence type diversity decreases. However, by 2009 sequence type diversity has returned to pre-vaccination levels. Although I showed that most sequence types associated with vaccine serotypes are absent from the population by 2007, I also demonstrated that a number of STs previously associated only with vaccine serotypes remain in the population, though now associated with a non-vaccine serotype.
8.2 Implications for disease prevention and suggestions for future work

One important aim of research into the evolution of *S. pneumoniae* and other bacterial pathogens is to inform treatment and prevention of disease. While I introduced and concluded each chapter by contextualising it in current trends in treatment and prevention where appropriate, in this section I will interpret broader conclusions from the thesis in the context of treatment and prevention.

Proponents of pneumococcal conjugate vaccination have suggested that in addition to reducing morbidity and mortality due to pneumococcal invasive disease, serotype specific vaccination could reduce the prevalence of antibiotic resistance (318,319). They argued that as drug resistance and multi-drug resistant strains are strongly associated with certain serotypes, particularly those covered by the conjugate vaccine, targeting these serotypes could reduce the incidence of antibiotic resistance (1,2). In Chapters 3 and 4 of this thesis I examined why certain serotypes were associated with antibiotic resistance but other serotypes were not and in Chapter 7 I examined the impact of vaccination on genetic diversity in a sample of carried pneumococci from a population of children. Given the conclusions of this research, I will discuss whether a decrease in the prevalence of antibiotic resistance is a logical outcome of serotype specific vaccination.

As I discussed in Chapter 7, the results from a study of longitudinally sampled, carried pneumococci, showed that the prevalence of total carriage did not decrease after vaccination was introduced (47). However, in the wake of vaccination, the prevalence of vaccine serotypes declined, while the prevalence of non-vaccine serotypes increased in carriage (47,309). This phenomenon (serotype replacement) has also been documented in surveillance of invasive disease (49,41). Even though the proportion of cases caused by non-vaccine serotypes has increased in those regions that have introduced vaccination, the clinical impact of this proportional increase is less clear (and beyond the scope of this research) (41). Thus, vaccination could only decrease antibiotic resistance if the replacing serotypes were less likely to carry antibiotic resistance. In Chapter 4 I showed that serotypes with a short duration of carriage were less likely to carry antibiotic resistance. I demonstrated that this relationship was preserved after the introduction of vaccination and there was no change in the prevalence of resistance in the sample of pneumococcal isolates from IPD. This outcome is expected as some serotypes that replaced vaccine serotypes in carriage and disease such as 19A have a long duration of carriage and as predicted by the results from Chapters 3 and 4, have a high prevalence of antibiotic resistance. Therefore, the introduction of vaccination will not necessarily reduce the prevalence of antibiotic resistance in carriage or disease if the replacing serotypes have long durations of carriage. However, if replacing serotypes are less likely to cause invasive disease, then the incidence of antibiotic resistant IPD could decline. The impact could be
particularly important in serotypes that are associated with meningitis (37), where penicillin resistance is clinically costly (as I outlined in Chapter 2). Thus, the impact of vaccination on antibiotic resistance is a topic deserving further study.

Antibiotic resistance remains present in the pneumococcal population and as I showed in Chapter 1 negatively impacts on our ability to treat meningitis. Therefore interventions to reduce the prevalence of antibiotic resistance in the community, such as restricting the use of antibiotics in the community, must still be pursued. The goals of such interventions are to prevent an increase in the prevalence of antibiotic resistance, prevent new resistant strains emerging and ultimately reducing the prevalence of antibiotic resistant strains (73,321,205,322). The logic of such interventions is that by removing selective pressure from antibiotic consumption, there is no selective advantage to resistant strains. If the resistant strain carries a fitness disadvantage, then, if selective pressure is removed from the population, the prevalence of resistant strains could fall. Therefore as the cost of resistance will determine the rate of decline in the prevalence of antibiotic resistant strains, estimating the fitness cost of resistance is an important part of addressing the possible impact of public health interventions to restrict antibiotic use.

Analysing the results from the mathematical model presented in Chapter 3, I showed that there was a consistent association between the prevalence of resistance, antibiotic use in the community and the threshold duration of carriage above which antibiotic resistant strains are maintained in the population. In Chapter 4 I began to develop and refine this method, through more sophisticated simulations and estimated an approximate fitness cost of antibiotic resistance. However, more sophisticated analysis and simulations are needed to refine these estimates. In addition analysing antibiotic resistance prevalence data from different settings with very low and high antibiotic use (such as Finland and Spain) would allow us to develop a more robust estimation of the cost of resistance.

In Chapters 5 and 6 I introduced selective sweeps and explored their impact in a generalised bacterial population. The results from these models predict that when an advantageous allele arises in a population it will sweep to fixation in that bacterial population. Of course this model was concerned with a very simplistic scenario and when an advantageous allele (or set of alleles) like antibiotic resistance arises in the pneumococcal population, the impact of selection has not been as predictable. In Chapters 3 and 4 I explored why certain serotypes did not develop antibiotic resistance as a consequence of the associated cost of resistance. However, the question remains as to why, within these serotypes, sensitive strains persist in the population. Some attempt has been made to resolve this question and reconcile the empirical evidence of coexistence with the
predictions of the simple model of selection and transmission (84). However, the results from the model presented by Colijn et al. allowed coexistence only under a small range of parameters (85), thus further work is needed to explore the phenomenon of coexistence in pneumococcal serotypes. Ultimately the answer to this question is likely to be found using a synthesis of parameter estimation from natural populations of bacteria, mathematical models to simulate and test hypotheses and large datasets, preferably collected from carriage.

8.3 Concluding remarks

In conclusion the research in this thesis has demonstrated that if treated with penicillin, a patient infected with penicillin non susceptible meningitis is twice as likely to die as a patient with penicillin susceptible meningitis. This demonstrated that understanding how and why antimicrobial resistance occurs in the pneumococcal population has a clinical impetus. I showed that the development of resistance is not an inevitable result of antibiotic use. As there is a cost of resistance, antibiotic resistance is unlikely to persist in serotypes which have a short duration of carriage. I demonstrated that rates of recombination in bacterial populations are not sufficiently high to prevent the loss of genetic diversity as a consequence of the selective sweep of a novel genotype carrying an advantageous allele. Finally I showed that the impact of vaccination can be detected by examining changes in the population genetic structure of \textit{S. pneumoniae} in carriage.

Throughout this research I have shown that mathematical models can be used to develop an understanding of a system and develop and formalise hypotheses. These hypotheses can then be examined in data from natural populations of bacteria. Finally, I have demonstrated that to understand the impact of treatment and prevention of disease, it is essential to understand the population genetics of an organism.
Bibliography


Appendix A: Glossary of laboratory techniques

**Acetyl Co-A synthetase** Acetyl Co-A synthetase is an enzyme involved in metabolism of carbon sugars. It is in the ligase class of enzymes, meaning that it catalyzes the formation of a new chemical bond between two large molecules (1).

**Cell density** Cell density can be determined using optical techniques based on the absorbance of light by cells using a spectrophotometer or colorimeter (2). With cultures well dispersed it is found that optical density remains proportional to cell density throughout the positive growth stages of cultures (3).

**Cell mass** Both viable and total cell counts do not measure increases in cell mass during the cell division cycle. Growing bacterial cells increase their mass before division; therefore even though the number of cells may not have increased growth is occurring. Cell mass can be determined directly by measuring cell weight after they have been isolated and dried.

**Chemostat** A chemostat is a continuous culture device used for growing and harvesting microbes. Chemostats are open systems where media is continually added and spent media and cells removed. The media is constructed so that one of the nutrients in the media is added in amounts that limit the number of bacteria. As media is constantly added and waste removed, limiting conditions do not occur and so growth never slows (4). The chemostat consists of three primary parts, the nutrient reservoir, the growth chamber and an efflux reservoir for exhausted medium, living cells and cellular debris (5). The most important feature of the chemostat is that all fermentation parameters such as growth chamber volume, nutrient concentrations and pH remain constant throughout an experiment (6). The rate of cell division can be manipulated by varying the input of nutrient from the nutrient reservoir until the desired cell density is reached. The equilibrium cell density is maintained as the number of new cells created by division is balanced by the number of cells siphoned off in the outflow. The equilibrium cell density is determined by the concentration limiting nutrient in the reservoir (4).

An assumption of the chemostat and much of the analytical and computational work attached to it is that there is a constant relationship between the amount of limiting nutrient inputted into the system and the population size. This leads to the further assumption that the population size can be maintained constant throughout the experiment. The assumption that reproduction is proportional
to nutrient uptake is a simplification and the evidence for this assumption is based on experimental
evidence rather than resting on any physiological basis (5).

**Fitness assays** The fitness of a bacterial type is measured relative to its ancestor. Relative fitness is
measured by allowing an evolved type to compete with its ancestor. The two types are grown
separately in the competition environment in order to ensure that they are equally acclimatised to
test conditions. They are then mixed, usually at the ratio 1:1 and grown in the competition
environment. Initial and final numbers of cells are assessed by the pour plate method [defined
below] to estimate a count of viable cells using an agar containing an indicator substrate. The growth
rate of each competitor is calculated as the natural logarithm of the ratio of its final density to its
initial density. Relative fitness is the ratio of the growth rates of the evolved and ancestral types (6).

gap A is a gene which codes for an enzyme in the glycolysis pathway (1).

*Lac operon* encodes proteins which utilize lactose and other β-galactosides (1).

*lacY* is a gene which codes for the LacY protein responsible for the membrane transport of lactose
and other β-galactosides (1).

*malT* is the gene which encodes activator proteins in the *mal* gene system, which codes for proteins
necessary to utilize maltose a type of sugar (1).

**Metabolite measures** Growth can also be estimated indirectly by monitoring the consumption of a
substrate by the bacteria. This techniques measure a result of growth rather than cell number.
Assays that measure the depletion of a substrate like glucose are used, as are those that monitor the
utilization of oxygen during respiration (9).

*mgl locus* is a cluster of genes which code for proteins involved in the uptake and utilization of
galactose which is a type sugar (1).

*mlc* is a gene which codes for a global repressor affecting sugar transport genes (1).

*pho A* is a gene on the Pho regulon which codes for an enzyme in the pathway of phosphate
assimilation from environmental organophosphate (1).

*Pho regulon* is a system of genes which code for proteins necessary to uptake and utilise phosphorus
sources for growth (1).

**Serial transfers of batch culture** An inoculum of non-growing bacterial cells is added to fresh
medium. A lag period occurs initially, where there is little or no growth in the culture as cells adapt
to the new culture conditions. Shortly after growth starts, cells begin to grow exponentially, at a maximum growth rate until some component of the medium becomes limiting. When nutrients become exhausted or toxic waste products build up bacterial growth is inhibited (2). The cells then enter the stationary phase where the growth rate is null (4). The progress of bacteria through the growth stages is monitored through the calculation of changes in cell density; by using optical cell density techniques and impedance counting or calculating cell concentrations determined by performing indirect viable cell counts. When the bacteria reach stationary phase, the inoculum of bacteria for the next flask of fresh medium and this can be repeated for the desired number of generations (4).

**Techniques to monitor the change in the prevalence of a neutral marker** Periodic selection was first observed when the frequencies of mutants carrying neutral markers observed in culture were altered by selection for an unrelated and unknown mutation at another locus (7,5). The sweep of the advantageous mutant types through the population are tracked by monitoring the fluctuations of a neutral marker in the population.

The densities of the different types are estimated by performing viable cell counts using an indicator agar which identifies the neutral marker by sight alone. For example, if the neutral marker on the wildtype cell was T5 phage resistance, one plate of culture medium would be inoculated with T5 phage so that only those containing the neutral marker could grow. A phenotype that is selectively neutral, such as T5 phage resistance, when phage is not present, should increase linearly in the population as mutations to generate this phenotype occur at a constant rate. Compared to the size of the population the T5 resistant phenotype is rare, so it is unlikely that an advantageous mutation will occur in a cell with this phenotype. A fall in the frequency of T5 resistant cells in the population can be interpreted as an increase in the frequency of an advantageous mutant (6).

**Total cell count** The total cell count is the number of cells, living and dead, that are present in a sample. This can be estimated by using a direct cell count which measures all the cells present. Cell counts can be measured using a counting chamber, where cells are directly observed under the microscope or using impedance counting, where cells passing through a narrow chamber disrupting the flow of an electrical current (8).

**Viable cell count** Using the pour plate method, an inoculum of cells in a suitable dilution is mixed with tubes of melted growth medium. The mixture is poured into a petri dish and the number of colonies counted is taken as the number of viable cells in the original inoculum of culture medium. An alternative method is the spread plate method where a small amount of cells in culture medium
is spread over the surface of a solid plate of growth medium. In both methods the plates are incubated and the number of colonies formed on the plate counted (9). One limitation of this method is that it assumes that each colony is descended from one cell. Another is that colonies must be visible to the naked eye before they are counted so a bacterium could have produced hundreds of thousands of descendants and still be classed as unviable cell by this method and therefore not counted by a viable cell count method.

References