Mathematical and statistical modelling of Sexually Transmitted Infections and their impacts on fertility

Torrington Callan

A thesis presented for the degree of

Doctor of Philosophy

Faculty of Science University of Technology Sydney

Australia

August 2021

Declaration

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Torrington Callan declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: June 2022

Abstract

Adverse reproductive health outcomes, such as pelvic inflammatory disease, ectopic pregnancy and tubal factor infertility, have been associated with *Chlamydia trachomatis* and *Neisseria gonorrhoea* infections.These reproductive health outcomes could be complemented by measuring subsequent pregnancies to assess impact on fertility. In this thesis, a number of statistical and mathematical modelling approaches were developed and utilised. Throughout the collection of results, the value of using multiple modelling techniques on a similar scientific domain is demonstrated by elucidating the reproductive burden of *Chlamydia Trachomatis* and the host and pathogen factors that moderate this. Using a retrospective data linkage cohort study of 132,962 women based in QLD between 2000 and 2005, we found that women with a history of testing for Chlamydia and/or Gonorrhoea had reduced odds of pregnancy, depending on the age of the women at testing and any prior occurrences of pregnancy (aOR 0.91, 95% CI 0.87 - 0.95 for women aged 20 with no prior pregnancy compared to no history of testing, aOR 0.71, 95% CI 0.68 - 0.75 for women aged 25, aOR 0.87, 95% CI 0.79 - 0.95 for women aged 30). After adjusting for Indigenous status, using a multiple imputation process to account for cases where women identified differently in different data sources, we found that non-Indigenous women with multiple positive tests were significantly less likely to be pregnant than women with all negative tests, whilst there was a small reduction in pregnancy odds following a single positive test (aOR 0.66, 95% CI 0.44 - 0.99; aOR 0.93, 95% CI 0.83 - 1.04 respectively). The same results were not replicated for Indigenous women (aOR 0.92, 95% CI 0.53 - 1.60, aOR 1.34, 95% CI 1.08 - 1.65 for multiple positive and single positive tests respectively), suggesting a set of demographic and biological factors moderated these results. The progression to pathology is thought to be

governed by ascension of *Chlamydia Trachomatis* to the upper genital tract. We developed a mathematical model based on a set of stochastic dynamics to investigate the moderating impacts of ascension. We found that ascension was dependent on the neutrophil response, and that the interplay between the neutrophil response, menstrual cycle timing and initial infectious load were critical factors in moderating ascension. The distribution of the number of bacteria ascending had a long right-tail (of all simulations, 36% showed bacteria ascending, but only 9% showed 1000 or more bacteria ascending, and 0.1% showed more than 10,000 bacteria ascending).

Acknowledgements

I'd like to acknowledge the significant contribution of my supervisor's for the work in this thesis and for their guidance in the start of my research career. James Brown, for his help and education in how to understand and interpret the statistical results we found in the data, and his effort and feedback to put these results into a sensible format. Stephen Woodcock, for his assistance and many long discussions on how to translate a rough understanding of biology into a precise mathematical model, and his help in understanding these results. Wilhelmina Huston, for her efforts and work in assisting the publication of these results and, for her guidance and education on the science and biology of what we were studying. I'd like to thank our collaborators, Joseph DeBattista, Brooke Berry for their work in collecting the data and interpretation of key results for the work presented in Chapters 2 and 3. Jane Hocking for her help in interpreting results, and guiding the presentation of results, particularly for the study published that forms a basis for Chapter 2. On a personal note, I'm very appreciative of the team at Spriggy for their help in making me a more knowledgable and more articulate practitioner, and for giving me the opportunity to learn and to still work on this thesis. I'd also like to note the contribution of the Salty Paddlers kayak squad, for keeping me happy, strong and healthy over the last few years. Most importantly, I want to note the contributions of my partner, Rachel and my son Louis, for their part in creating a life more enjoyable and fulfilling.

Contents

List of Figures

List of Tables

List of Publications

[1] **Callan, Torrington** and Stephen Woodcock. "Stochastic Modelling of Chlamydial Infections". In: *ANZIAM J.* 61 (2020), pp. 89–103. DOI: https://doi.org/10.21914/anziamj.v61i0.15159.

[2] **Callan, Torrington**, Stephen Woodcock, and Wilhelmina M Huston. "Ascension of Chlamydia is moderated by uterine peristalsis and the neutrophil response to infection". In: *PLOS Computational Biology* (2021). Accepted.

[3] **Callan, Torrington** et al. "A retrospective cohort study examining STI testing and perinatal records demonstrates reproductive health burden of chlamydia and gonorrhea". In: *Pathogens and Disease* 78.6 (Sept. 2020). ISSN: 2049-632X. DOI: 10.1093/femspd/ftaa052.

1

Introduction

1.1 Introduction

In the scientific process, hypotheses about a particular process can be evaluated using a set of observations. These observations are interpreted through some form of mental model. Models are used to represent a current state of understanding about nature. A current state of understanding about a particular element of nature will be incomplete, and so there will be some uncertainty associated with the process of interest - what we refer to as a random process. Using abstractions of nature in a rigorous format allows for an investigator to understand the key properties of a process of interest, and allows for the prediction of new dynamics and phenomena.

In this thesis, models are viewed as the tools through which scientific hypotheses about a key process in nature are evaluated. A critical difficulty in this ideal is that scientific hypotheses do not map cleanly to processes, which do not map cleanly to the models used to investigate them.

Creating a model of a process of scientific interest can be divided into mathematical modelling and statistical modelling. In a mathematical modelling exercise, a current state of knowledge (reflected in the scientific literature) represented in a mathematical form. This form is usually a set of differential (or difference equations). Depending on the complexity of the mathematical model, implications the model for new observations can be derived using analytical or numerical tools.

Key quantities of interest may include the number and location of steady state solutions and the relative impact of variables with the equations. The development of a mathematical model from a body of published scientific literature allows for the aggregation of disparate sets of information into one representation.

Statistical modelling begins with a set of data that has been collated into a single location. The process under study may or may not be well understood, and so a model is created to capture features of the available data, and in turn the model is used to investigate hypotheses about the process of interest. As with mathematical models, statistical models can be to determine the relative influence of variables, including the assertion that a variable previously thought to be influential has no observable effect. The distinction between statistical and mathematical approaches is that the former is usually based in the phenomenological observations, rather than a mechanistic understanding of a process in the latter.

Common approaches in modern analytics, using machine learning techniques, utilise large sets of data to estimate complicated and high-dimensional relationships. Many of these semi and non-parametric approaches are often used in statistical models as well. In the statistical models presented in Chapter 2 and 3, splines are utilised to estimate nonlinear relationships between age, exposure and outcome variables. These approaches require a significant quantity of data, which could be improved with a fully parameterised functional form, as is common in mathematical modelling approaches.

The key methodological hypothesis of this thesis is that a combination of statistical and mathematical modelling approaches will provide greater understanding when applied to a set of related research problems, and in a single applied problem. Mathematical models are more informative when confronted with data, by utilising common statistical methods, and statistical models are more relevant when the associations estimated can provide a mechanistic interpretation.

1.2 Background to the applied problem

In this thesis, the combined usage of mathematical and statistical modelling approaches is applied to a particular research problem, in order to fully demonstrate the value of combining both approaches. The particular research problem investigated is on the associated reproductive burden of two sexually transmitted bacterial pathogens, *Chlamydia Trachomatis* and *Neisseria gonorrhoeae*, and how the reproductive burden may vary between demographic, behavioural and host factors that impact bacterial replication, transportation, transmission along with confounding factors that also impact reproduction. A relevant summary of the literature regarding these STIs is presented in the introduction to each chapter separately.

1.3 Structure of this thesis

In Chapter 2, we use statistical modelling approaches on a large observational dataset to investigate the reproductive burden associated with testing and infection with the bacterial pathogens Chlamydia and Gonorrhea. The structure of the statistical models are derived from some principles on the current state of knowledge about the problem. We further investigate how the structure of the statistical models, including splines for continuous nonlinearties, interactions for effects that differ between groups, and hierarchical models for effects that vary across a large number of groups can influence the inference made of the process of interest.

In Chapter 3, we investigate a confounding effect of Indigenous status in the data used for Chapter 2. Deeper statistical approaches using multiple imputation are applied in order to elucidate more information of the process being studied, and we show how this approach furthers our understanding of the reproductive burden investigated in Chapter 2.

In Chapter 4, a simple mathematical model of birth weight conditional on gestational age is derived from a prior understanding of the process of pregnancy. A statistical model is then developed to evaluate the impact of Assisted Reproductive Technologies (ART) and testing history on the pregnancy process. We show that

combining a mathematical model with statistical estimation procedures can further our scientific understanding over a more simplistic statistical model.

In Chapter 5, a stochastic model of the basic within-host dynamics of chlamydia is developed. We discuss how the inclusion of randomness influences our observed outcomes, and use simulations from the model to investigate the global properties of infection dynamics within a host.

In Chapter 6, the basic stochastic model is extended to investigate the key moderating host and pathogen factors of the progression of an infection to pathology.

Overall this thesis will utilise a number of statistical and mathematical modelling approaches. The aim will be to demonstrate that employing a number of tools on a single scientific problem will lead to a deeper understanding of the process of nature being studied.

2

A retrospective cohort study examining STI testing and perinatal records demonstrates reproductive health burden of chlamydia and gonorrhea

2.1 Introduction

Chlamydia trachomatis (chlamydia) and Neisseria gonorrhoeae (gonorrhea) are commonly diagnosed sexually transmitted infections (STI) that have been associated with serious reproductive health outcomes for women, and may cause obstetric or perinatal complications [22, 19].

The reproductive burden of chlamydia and gonorrhea is most often evaluated using Pelvic Inflammatory Disease (PID), although there are a number of other causes of PID outside of these two pathogens. A cohort study based in Uppsala County, Sweden, established an association between chlamydial infection and the incidence of PID [21], by using a cohort of women with positive and negative tests for chlamydia, and comparing to a reference group of women who were never tested. Notably, the study found a significant increase in the incidence of PID following a positive test, when compared to a negative test (5.6% (95% CI 4.7-6.7) in women with positive tests versus 4.0% (95% CI 3.7-4.4) for women with negative

tests. The study also found that screening was also a risk factor for PID (incidence was 2.9% (95% CI 3.6 - 4.4) for those never tested), after adjusting for education status, age, income and housing. The relationship between chlamydia testing and infection was further investigated in a population-based retrospective cohort study in Denmark of 500000 women [11]. This study also found that positive tests for chlamydia increases the risk of PID compared to only negative tests, and that any history of testing increased the risk of PID compared to the absence of any testing history.

A population based cohort study of 315,123 women in Western Australia established a link between PID and [24] testing for chlamydia and/or gonorrhea, after adjusting for age, Indigenous status, year of follow up, health area and socioeconomic status. The study found that gonorrhea infection substantially increased the risk of PID compared to chlamydia. Other population level assessments have robustly established chlamydia to be a significant risk factor for PID, even for some studies where test results have been negative [8, 9, 10, 25, 27]. However, PID is not a notifiable condition, diagnosis is subjective and may be inconsistent between different settings [12].

The Prevention of Pelvic Infection trial was a randomised control trial of young, sexually active women, that provided study participants with self-administered vaginal swabs and a questionnaire [23]. Samples were randomly allocated to a treatment group, where the sample was immediately tested for chlamydial infection and treatment was provided if necessary, or a control group where the analysis of the sample was deferred by 12 months. The trial found an increased incidence in PID for women in the deferred screening controls (7/74 samples or 9.5% that had tested positive for chlamydia, compared to the treatment group with an infection (1/63 samples or 1.6%). This evidence suggested that an increase in incidence of chlamydial infection also increases the incidence of PID. However, a majority of the episodes of PID occurred in women with a negative test for chlamydia (30/38 diagnoses in the study, or 79%).

An analysis of data collected by England's National Chlamydia Screening Programme (NCSP) was compared to national survey data, to assess the differences

in demographics and behaviour between people accessing testing services, and testing positive for chlamydia [28]. The analysis found that women accessing testing services were more likely to have had 2 or more sexual partners in the last year 47% vs 30%), and less likely to have used a condom in their most previous sexual encounter (32% vs 37%). These factors, independent of data source and adjusted for age and ethnicity, were increased the chances of chlamydia positivity.

Pelvic Inflammatory Disease (PID) has also been associated with increases in the risk of ectopic pregnancy and tubal factor infertility. A study of 2,501 women with clinical risk of PID using diagnostic laparoscopy found 1,844 abnormal findings [32]. Of these women, 1,732 with abnormal findings and 652 with normal findings (control subjects) were followed, and it was found that 16% of women with abnormal findings could not conceive (after attempting) and 10.8% had confirmed tubal factor infertility, compared with 2.7% and 0% of the control subjects respectively. Later studies further confirmed a link between PID and the risk of ectopic pregnancy and tubal factor infertility [18, 21, 29, 6, 23, 8, 9, 24, 27]. Tubal factor infertility and ectopic pregnancy have been associated with a history of both STIs in numerous studies [30, 18, 7, 5, 13, 26]. Reinfection events have been associated with additional increases in the risk of PID [9, 11], however the presence of multiple positive tests has not been shown to additionally increase the risk of tubal factor infertility (that is, the risk from one infection is statistically identical to that of multiple) [11]. Obstetric or perinatal complications associated with STIs during pregnancy include pre term deliveries, pre term rupture of membranes, low birth weight or still birth [20, 17, 24].

In spite of the substantial proof for these reproductive health burdens, there is limited evidence evaluating subsequent pregnancy, which has been substantiated with a birth record, as an indicator of fertility after chlamydia or gonorrhea infections. We are aware of only one previous study that has evaluated pregnancy (as a measure of fertility), substantiated by a birth record, and associated with a history of chlamydia testing. This was a study of a Norwegian population (n = 20, 762) that found that women with a recorded test for chlamydia that found a positive test had a slightly increased incidence of births, compared to negative tests

(Crude HR 1.07; 95% CI 1.01–1.12) [5]. Of 3,321 women with a positive test for chlamydia, 1659 had a pregnancy and 1662 did not. 17,441 had only negative tests for chlamydia, of whom 9169 had a pregnancy and 8272 did not. The derived crude odds ratio from this study, comparing positive and negative tests for substantiated birth record was therefore 0.90 (95% CI 0.84 - 0.97) (the odds ratio was derived from the numbers presented in the paper). Importantly, this study does not provide a comparison to an unexposed group who were not tested for chlamydia, which was shown in previous linkage studies of PID to be a significant risk factor [5].

A robust measure of pregnancy that has been substantiated by a birth record as an outcome measure would add to current understanding of the reproductive health burden from these STIs. We have used retrospective cohort study design to evaluate the association of pregnancy (via birth records) and perinatal outcomes (pre term or low birth weight and still birth) with chlamydia and/or gonorrhea testing history and test results in women in the State of Queensland, Australia. Based on previous studies of STI testing history and PID, we expect to see a reduced pregnancy rate for women exposed to testing and diagnosis of Chlamydia and Gonorrhea.

2.2 Methods

2.2.1 Study design and data retrieval

We used a cohort study to investigate associations with chlamydia and gonorrhea testing history and pregnancy outcomes (using retrospective data linkage methodologies). The datasets used are based in the State of Queensland, Australia ($\sim$20% of Australian population). Our primary outcome was pregnancy defined as a record of birth in the Queensland Perinatal Registry. We used the birth record to define pregnancy because this provided a substantiated measure that a pregnancy had occurred. Secondary outcomes included perinatal outcomes of low birth weight ($<$2500 g) and pre term birth (before 37 weeks). Secondary outcomes were derived from the Queensland Perinatal Registry, and so analysis for these outcomes was restricted to women with a birth record (pregnancy).

We compiled the cohort by merging two datasets that we obtained from AUSLAB, the Queensland Health Clinical and Scientific Information System. AUSLAB Records cover any testing conducted through a public provider, and so it does not necessarily include testing conducted through private systems. Our cohort included two groups of women: (1) an exposed group that comprised women who had had a chlamydia or gonorrhea test between 2000 and 2005; and (2) an unexposed group that comprised women who had a full blood count laboratory investigation but no chlamydia and/or gonorrhea test between 2000 and 2005. Records of tests for chlamydia and/or gonorrhea, conducted in all public pathology laboratories in QLD between January 1st, 2000 and December 31st 2005 for women aged between 15 and 31 were collated (Median age 24; IQR 21–26, at December 31st 2005). The unexposed group comprised women aged 15–31 years old (Median age 24; IQR 21–27, at December 31st 2005) who underwent a full blood count test (independent of indications for the test) in the same 5 year period and with no record of testing for chlamydia and/or gonorrhea during the study period (Figure 2.1). Entry into the exposed group was defined as having any testing history to chlamydia and/or gonorrhea, as women who access testing services have characteristics that are more likely to result in a positive diagnosis. Previous studies using PID as an outcome measure have also found that women with a testing history, but no positive diagnosis experience increased risk of PID [27]. Further analysis also was used to compare women in the exposed group by their diagnosis results, by classification into distinct groups based on testing status (negative or positive) and number of tests. Full blood count was selected as our unexposed because these were women who were also in contact with primary care and also did not have a test record in our data for chlamydia and/or gonorrhea. The FBC has a range of indications, but both FBC and STI screening can be used during antenatal or preparation for pregnancy investigations, therefore, analysis also included pregnancy during the testing window [31]. Deterministic methods were used to remove any women from the full blood count group that also appeared in the chlamydia and/or gonorrhea testing dataset.

We linked the testing dataset with the Perinatal Data Collection (Queensland Health Statistics Unit) that records all births in Queensland. The perinatal dataset included all birth entries from 01/01/2000 until 30/06/2013, which totalled

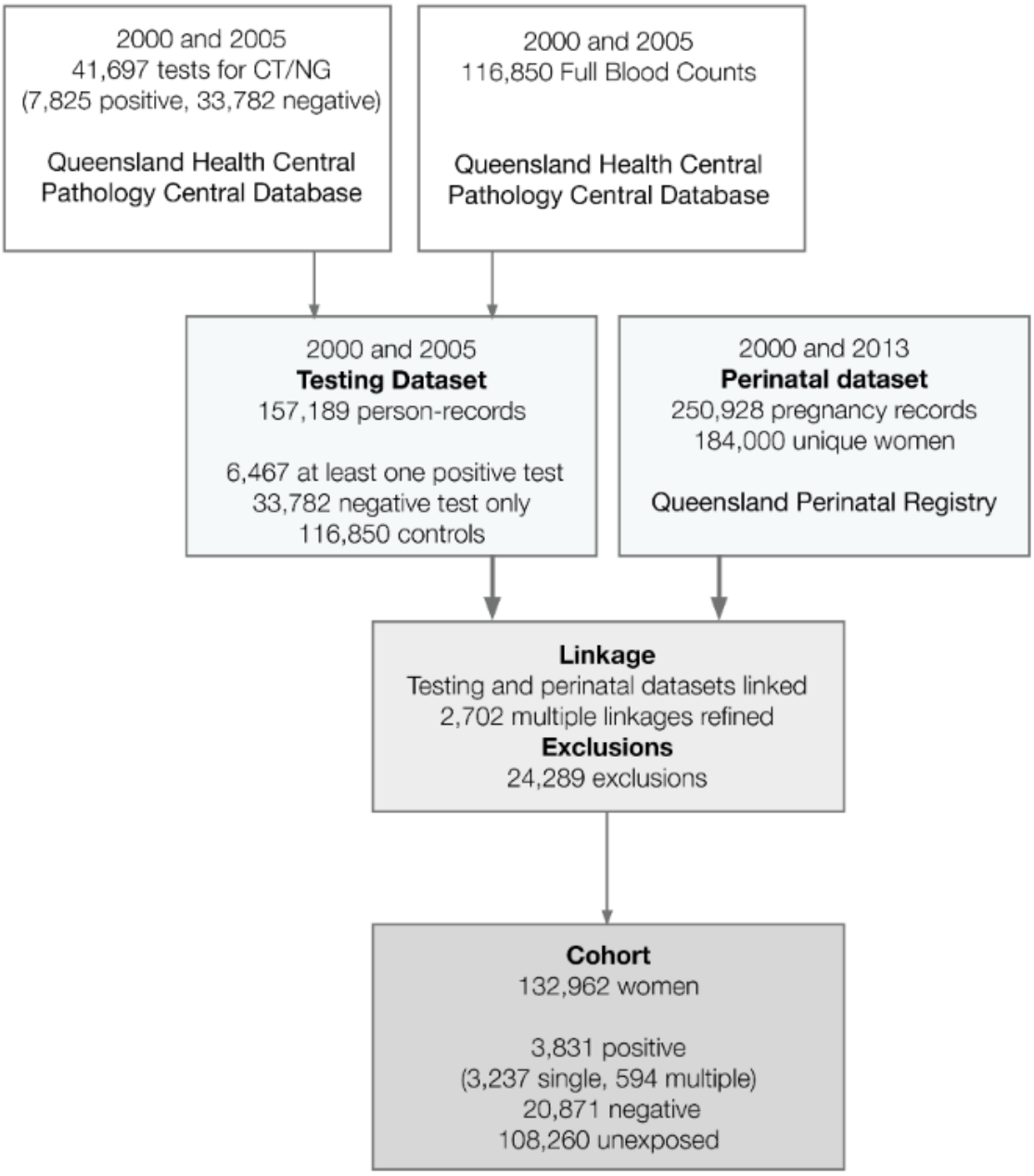

Figure 2.1: **Summary of cohort formation.** The cohort was formed using data linkage methodologies between a testing dataset and perinatal dataset as outlined in the figure.

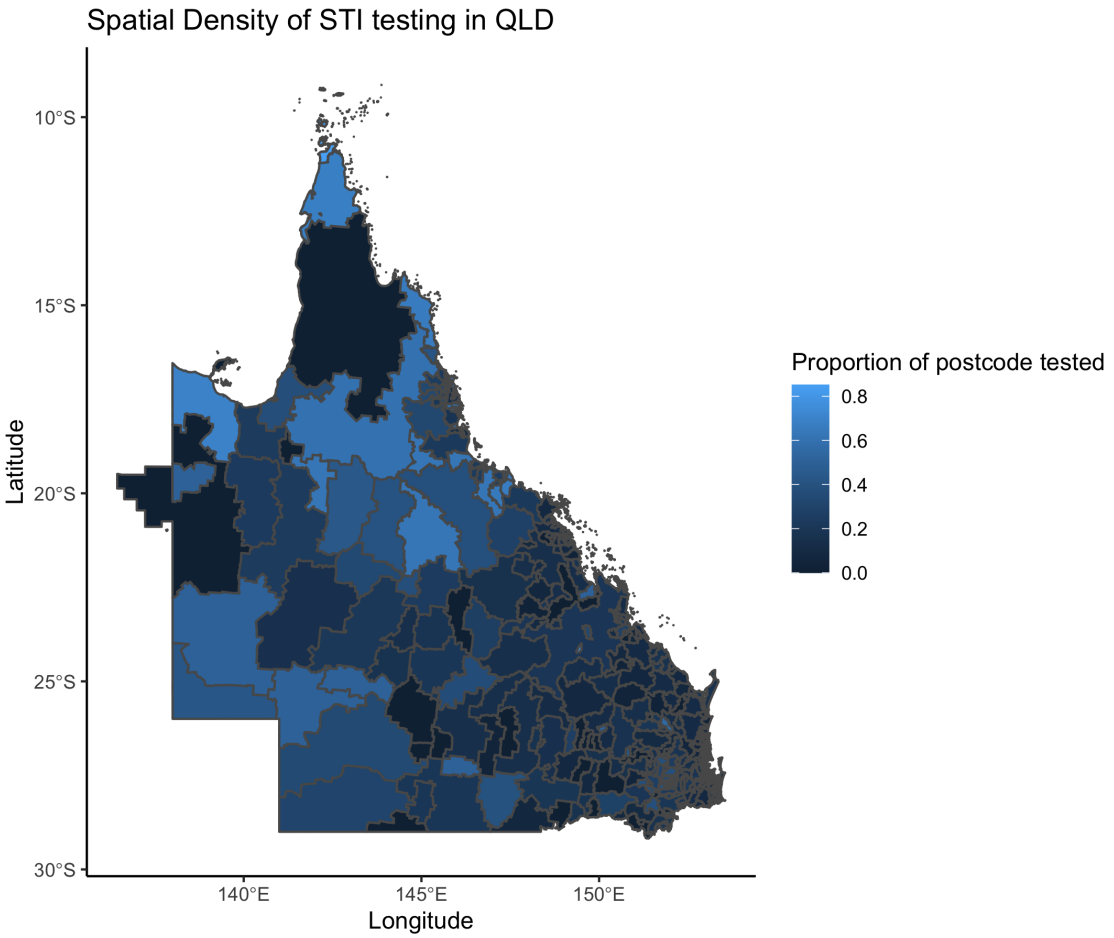

Figure 2.2: **Spatial map of testing data.** The figure shows the proportion of women in each postcode that were ever tested for an STI (in our dataset)

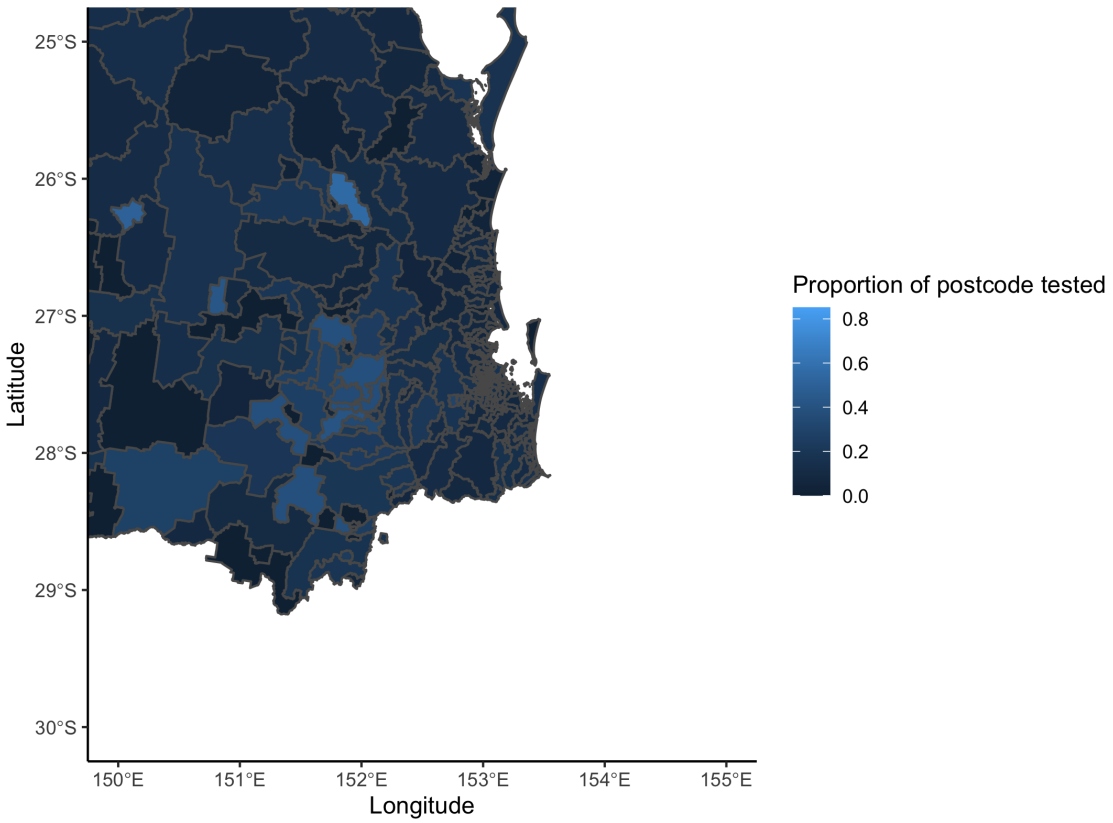

Figure 2.3: **Spatial map of testing data.** The figure shows the proportion of women in each postcode that were ever tested for an STI in the more densely populated South East region.

250,928 pregnancy records to 184,000 unique women. The perinatal data included the mother's date of birth, the baby's date of birth, birth weight, gestational weeks of the pregnancy and stillbirth.

2.2.2 Linkage methodologies

First name, middle name (if available), surname, sex and date of birth were used for linking the AUSLAB datasets with the Queensland Perinatal Registry. Linkage was deterministic in the case when information matched exactly, and probabilistic methods were used in the remaining cases. The linkage framework does not have a published error rate, but has a similar design to the Western Australia and New South Wales systems which have reported error rates of 0.11% and 0.3%, respectively (New South Wales Government 2012; Queensland Government 2017). Women whose data were unable to be linked with the Queensland Perinatal Registry were assumed to have no births during the study time frame. Multiple linkages (n = 2702) were randomly assigned a single linkage in the dataset (Fig. 2.4).

Women with missing postcode information (n = 19 764), or a postcode from outside of Queensland were removed (n = 4525). Linkages where made in these instances, but data with missing postcodes were removed as location and socioeconomic variables were key confounders in our study, and data with non QLD postcodes women who resided outside the state were less likely to have pregnancies recorded in QLD perinatal data (only 13.7% of women in this group had a recorded pregnancy). In order to test the impact of this assumption on our presented estimates, we ran the main effects model presented in the preliminary results section and found that the adjusted odds ratio for the exposed group was 0.86 (95% CI 0.84 - 0.90) and 0.91 (95% CI 0.89 - 0.93) for the prior pregnancy effect. This results are overstate the magnitude of the effect discussed in the results section. A total of 30 women without a date of birth were excluded. All women with a recorded birth that occurred before the test record during the 2000–2005 test were excluded. All person-records were assigned a de-identified code and the data were provided to the study investigators after linkage and de-identification, to protect the identities and privacy of the women whose personal information comprised the study.

Demonstration that there is not uncertainty due to linkage correction

In our data set, there are cases where the linkage between testing and perinatal datasets were not exact (probabilistic leading to more than one linkage). As a consequence, there are cases of a woman's testing records being linked to multiple perinatal records and vice versa. Our correction method takes only one of these linkages at random and deletes the rest.

We would like to demonstrate that this uncertainty induced by this method does not greatly impact the parameter estimates in our statistical models. To do this, we apply one iteration of the correction method (randomly deleting extraneous links) to obtain a complete dataset. Then we estimate our model of pregnancy, with the predictors of testing status (negative tested only, positive tested or never tested) and age as a continuous variable. From this model, we obtain the parameters of the log-odds for the negative tested group against the never tested, and for the positive tested group against the never tested group. These are converted into odds ratios. As a result, for each linkage error correction, we obtain a set of odds ratios that are of final interest in our study.

This procedure is then repeated 1000 times. That is, we randomly select a one-to-one set of linkages between datasets, estimate our model and obtain a set of odds ratios. We are then left with a set of 1000 odds ratios. The figure above plots the odds ratios for the negative and positive tested groups.

The figure was generated from this correction for linkage error. We can observe from the figure that the empirical uncertainty that results from the linkage correction method is small in comparison to the magnitude of the odds ratios. In our specific case, 95% of the estimates for the positive tested odds ratio were between 0.4671 and 0.4846, and 95% of the estimates for the negative tested odds ratio were between 0.3999 and 0.4063. By comparison, the 95% Confidence Interval for the positive tested group was (0.44, 0.50) and for the negative tested group the 95% CI was (0.39, 0.41), which represents the uncertainty from the model itself.

The law of total variance can be used to derive confidence bounds for these

parameters across all 1000 simulations. The law states that

$$Var(Y) = E\left[Var(Y|X)\right] + Var(E\left[Y|X\right])$$

, where X and Y are random variables on the same probability space. For the negative tested odds ratio, the total variance is $\frac{0.41-0.39}{3.92} + 0.0067$, and so the 95% confidence interval for this parameter including linkage uncertainty is (0.3825, 0.4275). Simi- larly, for the positive tested odds ratio, the total variance is $\frac{0.50-0.44}{3.92}$, and so the confidence interval for this parameter including linkage uncertainty is (0.4107, 0.5393).

We can state that error induced by the linkage correction is smaller than the uncertainty already represented in our models, and therefore this will have no impact on our interpretation of the final results. We use the completed dataset from one of the applications of the correction method for the analyses reported in this chapter.

2.2.3 Statistical methods

We use a series of logistic models estimated with maximum likelihood and Bayesian methods, that include spline components and random effects, building to a hierarchical generalised additive modelling to determine the association between chlamydia and/or gonorrhea testing history (and test results) and each outcome. Generalised additive models. are a way to extend logistic regression models to include terms that are nonlinear in flexible ways. Hierarchical models can be extended to situations where data is collected in clusters, rather than independently distributed. Information about the testing dates for women in the unexposed group, and women with only a negative test result was not available, which restricted any valid time-to-event analysis. Analysis assessed the odds of pregnancy in a discrete outcome period from 1/1/2007 until the end of the follow up period on 30/06/2013.

Entry into the cohort was based on testing records in AUSLAB (either chlamydia and/or gonorrhea test [exposed] or blood test [unexposed]) in 2000–2005. The cohort was followed up to assess for birth and secondary outcomes (birth weeks, birth weight) until 30/06/2013. The study analysed only the first recorded birth for

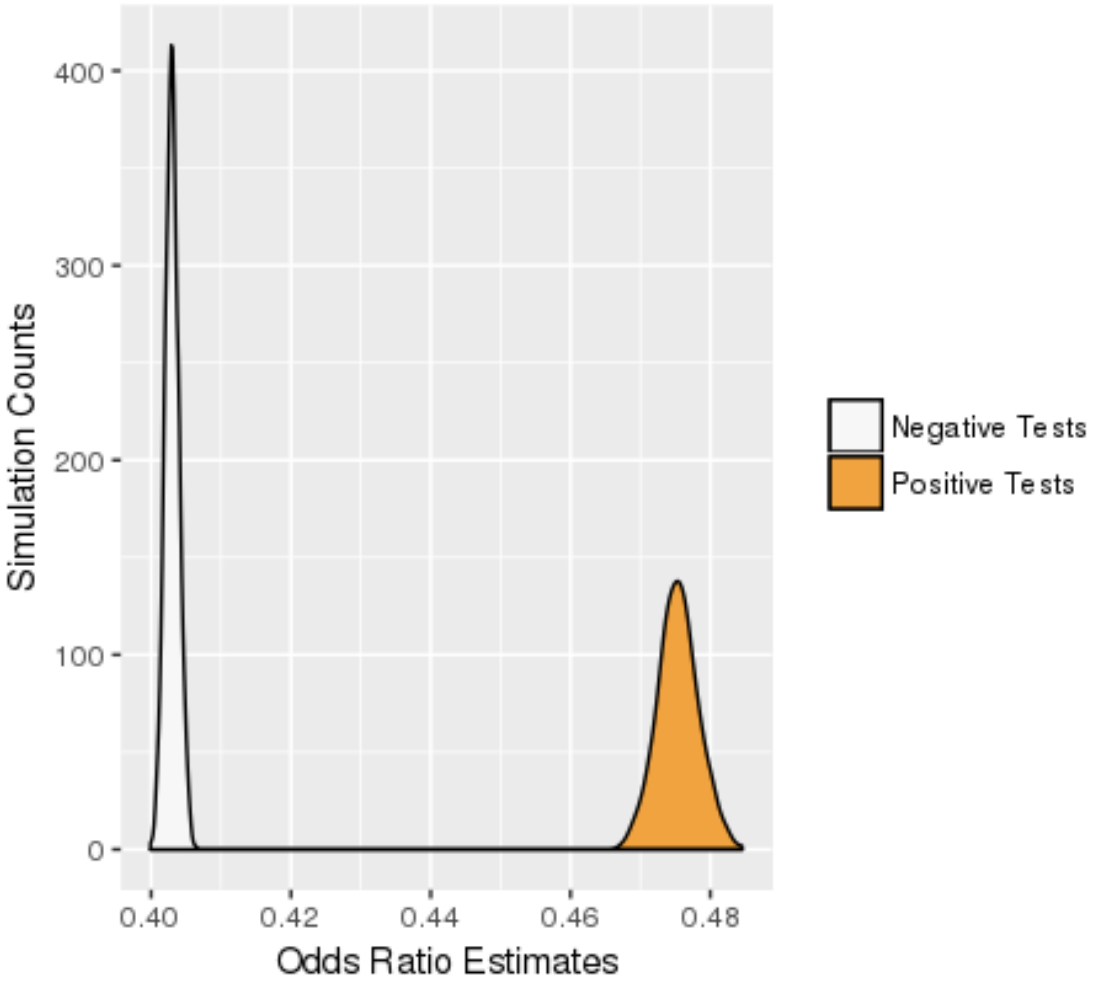

Figure 2.4: **Distribution of parameter estimates for 1000 iterations of the linkage error correction.** The figure shows the odds ratio ranges for all 1000 iterations of the random deletion process for the person records with multiple linkages, data is clustered by negative and positive chlamydia and/or gonorrhea results.

each woman. We converted birth weight (low birth weight as a birth weighing less than 2500 g) and birth weeks (a pre-term birth is one born before 37 weeks of gestation) to binary outcomes for the purposes of this study, using values from World Health Organisation standards 2006 [1]. These outcomes represent significantly increased risks of perinatal mortality, morbidity and longer term developmental issues [**hinz**]. Postcodes were used to classify the socio-economic area of the women, defined as low, middle or high using the ABS Index of Relative Socio-economic Disadvantage and to classify regionality as living in a major city, regional or remote geographical area [2]. If a woman had multiple testing events the first postcode recorded was used. Location was then defined as the combination of socioeconomic status and regionality, to account for the interaction effect between an area's socioeconomic status and overall access to health services.

Data preparation and analysis was completed using the R software version 3.6.3. The code used for this analysis can be found at github.com/torricallan/QLDPregnancyStudy.

Model selection

Our methods for specifying the models used in this analysis largely follow the recommendations of Harrell [15], which favours a first principles approach to model selection over stepwise methods that introduce bias into final estimates and avoid a deep analysis of the problem.

Models were adjusted for age, as the odds of pregnancy are highly age-specific and younger women in the cohort may have much different susceptibility to the development of pathology, and we are more likely to observe more of their reproductive lifespan in the study follow up time. We adjusted for pregnancy in the testing period, as this measured women with no obvious fertility problems having greater overall chances of becoming pregnant again. The last main adjustment variable was for socioeconomic status and location, which acted as a measure of access to health services [16]. and removed the confounding of exposure status with socioeconomic status on pregnancy. Age was interacted with pregnancy in the testing period and exposure status, in order to evaluate the impact of association as a function of follow-up time relative to age. Pregnancy in the testing period was

interacted with exposure status, as pregnant women were often tested for STIs as part of routine pregnancy checks. Age was modelled with a restricted cubic spline with three knots, as we believed that the effect of age on pregnancy was unlikely to be linear, and felt that limited data in the tails of our age distribution might influence the choice of polynomial degree. Three knots was chosen as it was the most parsimonious choice for the spline selection, whilst still being flexibly nonlinear. Perinatal models used the presence of the outcome (pre term birth or low birth weight) in the testing period as a predictor for the outcome in the outcome window, as we hypothesised that perinatal complications would be shared across pregnancies within women. Perinatal models were also adjusted for maternal age when she gave birth, as this adjusted for both effects in time (via age at 31/12/2005), and effects of age changing the tendency to experience perinatal and obstetric complications. This second age variable was modelled as linear, as any nonlinear term would be non-identifiable with the primary age variable.

The models run in this analysis, for the main pregnancy and perinatal results are as follows. In 2.3.2 we model pregnancy as an outcome using testing history as the key exposure, with only main effects included. We then model pregnancy, with the key interaction between testing history and prior pregnancy. In subsection 2.3.3, we modelled pregnancy as an outcome using testing history as the key exposure variable, with the full set of interactions specified. We then modelled pregnancy using testing status (negative, positive once, positive multiple). Finally, we modelled pregnancy broken down by pathogen (CT or NG). All pregnancy models in this section were of the form

$$
\begin{aligned}
y_i &\sim Binomial(n = 1, p_i) \\
log(\frac{p_1}{1 - p_i}) &= \beta_{intercept} + \beta_{testing}(testing)_i + \beta_{priorpregnancy}(prior\ pregnancy)_i \\
&+ \sum_{k=1}^{K} \gamma_k(age)_i + \beta_{location}(location)_i + \beta_{testing:prior\ pregnancy}(testing : prior\ pregnancy)_i \\
&+ \sum_{k=1}^{K} \gamma_k(age : testing)_i + \sum_{k=1}^{K} \gamma_k(age : prior\ pregnancy)_i
\end{aligned}
$$

For the hierarchical models in 2.3.4, we use a Bayesian model for pregnancy with testing as the key exposure variable. The age effect was modelled with thin-plate smoothing splines, as opposed to restricted cubic splines in the earlier section. In order to compare the sensitivity of the results to the choice of spline, the population-level effects model, estimated using MCMC methods is included. The prior on the standard deviation of the splines was a standard normal. The prior on the parameters in the model were also given N(0,1) priors, as a standard normal prior roughly states that 95% of the prior mass is between -2 and 2 on the log-odds scales, which in turn implies that most of the prior mass is between 0.1 and 0.9 on the probability scale. This reflects our belief that no combination of variables would cause someone to have very high and very low probabilities of pregnancy, since wider priors on the log-odds scale tends to place more prior mass on the boundaries of the probability scale. The full varying effects model is as follows

$$
\begin{aligned}
y_i &\sim Binomial(n = 1, p_i) \\
log(\frac{p_1}{1 - p_i}) &= \beta_0 + \beta_{postcode} + \beta_{1,postcode}(testing)_i + \beta_{2,postcode}(prior\ pregnancy)_i \\
&\quad + \beta_3(location)_i + \beta_{4,postcode}(testing : prior\ pregnancy)_i \\
&\quad + \sum_{k=1}^{K} \gamma_k(age)_i + \sum_{k=1}^{K} \gamma_k(age : testing)_i + \sum_{k=1}^{K} \gamma_k(age : prior\ pregnancy)_i
\end{aligned}
$$

$$
\begin{aligned}
\beta_3 &\sim N(0, 1) \\
\gamma_k &\sim N(0, \tau_2) \\
\tau &\sim N(0, 1) \\
\begin{bmatrix} \beta_{postcode} \\ \beta_{1,postcode} \\ \beta_{2,postcode} \\ \beta_{3,postcode} \end{bmatrix} &\sim MVNormal([\mathbf{0}], \mathbf{S}) \\
\mathbf{S} &= \begin{bmatrix} \sigma_1, 0, 0, 0 \\ 0, \sigma_2, 0, 0 \\ 0, 0, \sigma_3, 0 \\ 0, 0, 0, \sigma_1 \end{bmatrix} \mathbf{R} \begin{bmatrix} \sigma_1, 0, 0, 0 \\ 0, \sigma_2, 0, 0 \\ 0, 0, \sigma_3, 0 \\ 0, 0, 0, \sigma_1 \end{bmatrix} \\
(\sigma_1, \sigma_2, \sigma_3, \sigma_4) &\sim N(0, 1) \\
\mathbf{R} &\sim LKJ(1.0)
\end{aligned}
$$

The Discrete time models (in subsection 2.3.5) follow the above format, with the exclusion of the spline terms for age and the inclusion of a coefficient for each discrete year of entry into the data. In subsection 2.3.6, we model pre term birth with testing and testing status as key exposure variables.

2.2.4 Ethical approval

This study was reviewed and approved by the Metro North Hospital and Health Service Human Research Ethics Committee (approval number HREC/14/QPCH/85).

2.3 Results

This results section is structured as follows. We first describe a series of models that predicts pregnancy in the outcome window (2006-2013) given exposure status (testing history) using main effects only, linear interactions and then nonlinear interaction terms. Each model is described in full to allow the reader to compare effects between different models, to better understand the impact of the adjustments made to the model. Following is a set of models that break down exposure status into testing categories and infection categories, which allows comparison of the effect of specific testing outcomes (such as testing positive) to the overall exposure effect. Then presented is a series of hierarchical models estimated with Bayesian methods. Again, a series of models are presented, with fixed effects only, varying (or random) intercepts and varying (or random) slopes to compare the impact on estimates of model structure. The next section presents a set of discrete time models for a specific age group. The models include main effects for exposure and time, interaction effects of exposure and time, and an interaction between exposure and time where the exposure effect varies across postcodes. Finally, models are presented that use perinatal outcomes (pre-term birth and low birth weight) rather than pregnancy as the outcome variable, where exposure is broken down by testing status.

All models are adjusted for pregnancy in the testing window (2000–2005), age at 2005 and location. The postcode of residence at the first testing location is included in some models. The definition of exposure was then broken down into testing categories (multiple positive tests, single positive test, all negative tests or unexposed), and infection categories (chlamydia positive, gonorrhea positive, chlamydia and gonorrhea positive, all negative tests and unexposed).

2.3.1 Cohort Description

The cohort comprised 132 962 women after linkage, refinement and exclusions (Table 2.1). The women in the cohort were aged between 22 and 39 years at the study completion, June 30th 2013, (Median 31; IQR 28–34). Overall, 24 702 (18.6%) women were tested for chlamydia and/or gonorrhea (exposed), including,

	Unexposed (% of group)	Exposed (%)	Total
Total			
	108,260	24,702	132,962
Age (at 2005)			
	Mean = 24.3 (IQR 21.7 – 27.1)	Mean = 23.4 (IQR 20.8 – 25.9)	Mean = 24.1 (IQR 21.5 – 26.8)
Location			
Major City and High SES	39,728 (36.7)	5,905 (23.9)	45,633 (34.3)
Major City and Medium SES	17,416 (16.1)	2,547 (10.3)	19,963 (15.0)
Major City and Low SES	7,457 (6.9)	978 (4.0)	8,435 (6.3)
Regional and High SES	6,017 (5.6)	1,961 (7.9)	7,978 (6.0)
Regional and Medium SES	22,339 (20.6)	6,883 (27.9)	29,222 (22.0)
Regional and Low SES	12,222 (11.3)	3,446 (14.0)	15,668 (11.8)
Remote and High SES	122 (0.1)	13 (0.05)	135 (0.1)
Remote and Medium SES	1,602 (1.5)	789 (3.2)	2,391 (1.8)
Remote and Low SES	1,357 (1.3)	2,180 (8.8)	3,537 (2.7)
Pregnancy in the testing period			
Not Pregnant	75,217 (69.5)	20,250 (82.0)	95,467 (71.8)
Pregnant	33,043 (30.5)	4,452 (18.0)	37,495 (28.2)

Table 2.1: **Summary of the cohort.**

20 871 (15.7%) women who had only negative tests(s), 3237 (2.4%) who had a single positive test recorded and 594 (0.4%) who had multiple positive tests recorded. The unexposed group included 108 260 women. A total of 69 533 (52.3%) women in the cohort had a pregnancy during the study time frame (Fig. 2.8). A total of 37 495 (28.1%) women in the cohort had a pregnancy during the testing window (2000–2005) and 43 325 (32.5%) had a pregnancy during the outcome period (2006–2013).

Women with a history of testing were on average younger than their unexposed counterparts (Fig. 2.5, **??**), more likely to be from a more regional and lower SES postcode (Fig. 2.7), and were less likely to be pregnant in the testing window (Fig. 2.8), hence the usage of these variables in all models used for investigation (Table 2.1 shows a numerical summary of the cohort description).

Overall, 43,325 (32.58%) women recorded pregnancies in the outcome window, of which 35,179 (32.50% of unexposed group) were unexposed and 8146 (33.0% of exposed group).

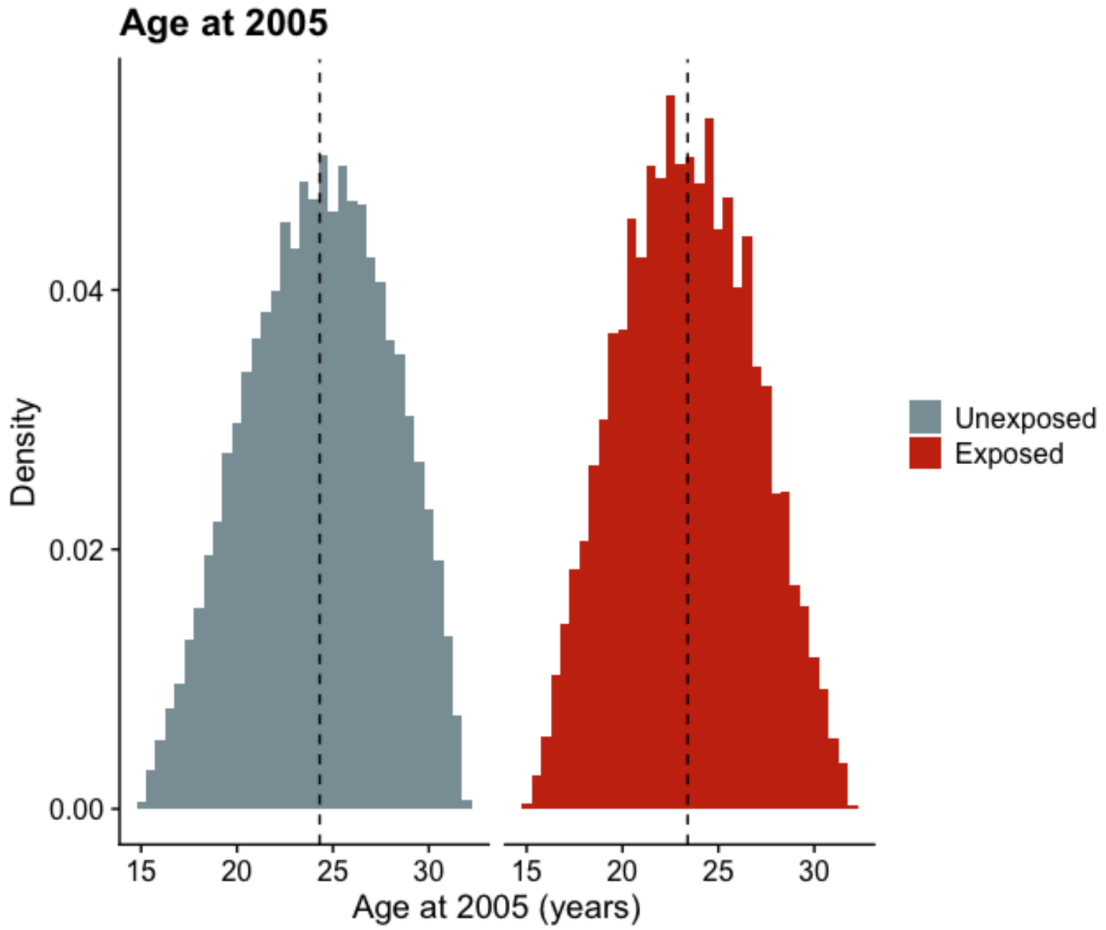

Figure 2.5: **The distribution of ages grouped by exposure (history of testing) group in the study cohort.**

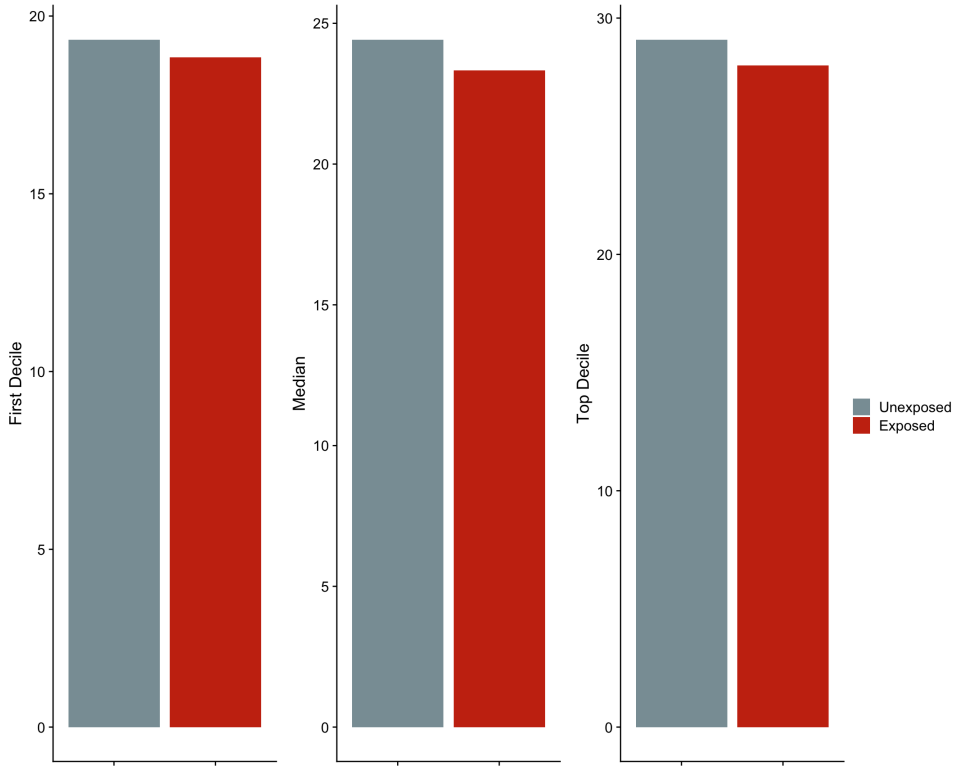

Figure 2.6: **Comparison of the 1st decile, median and 9th decile values for the age distribution for exposure groups.**

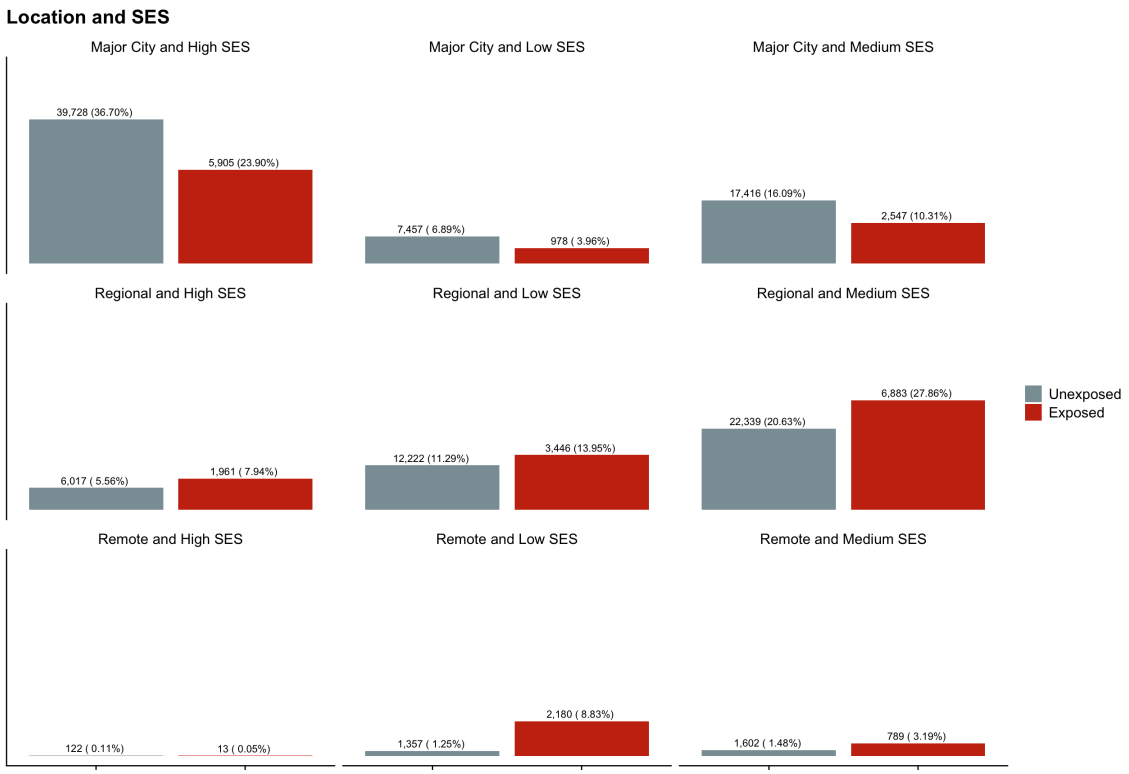

Figure 2.7: **Proportion of the cohort in each location and SES category by exposure.**

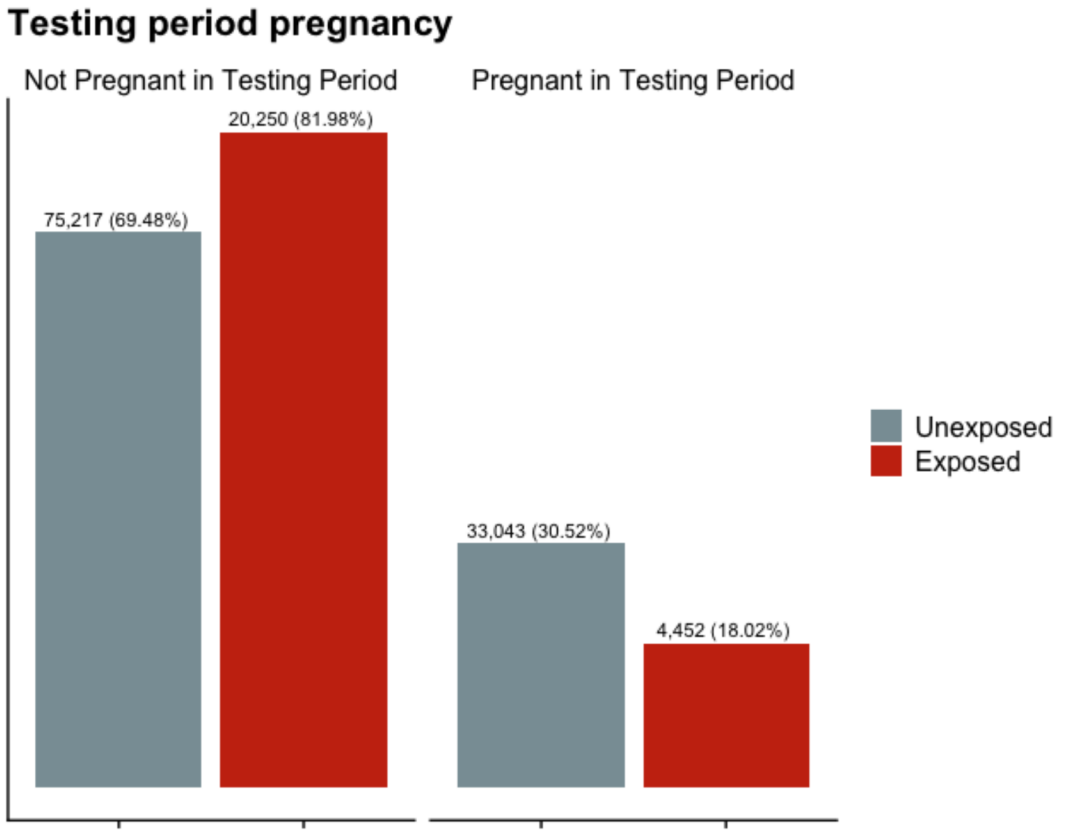

Figure 2.8: **Proportion of the cohort with a pregnancy in the testing period by exposure.**

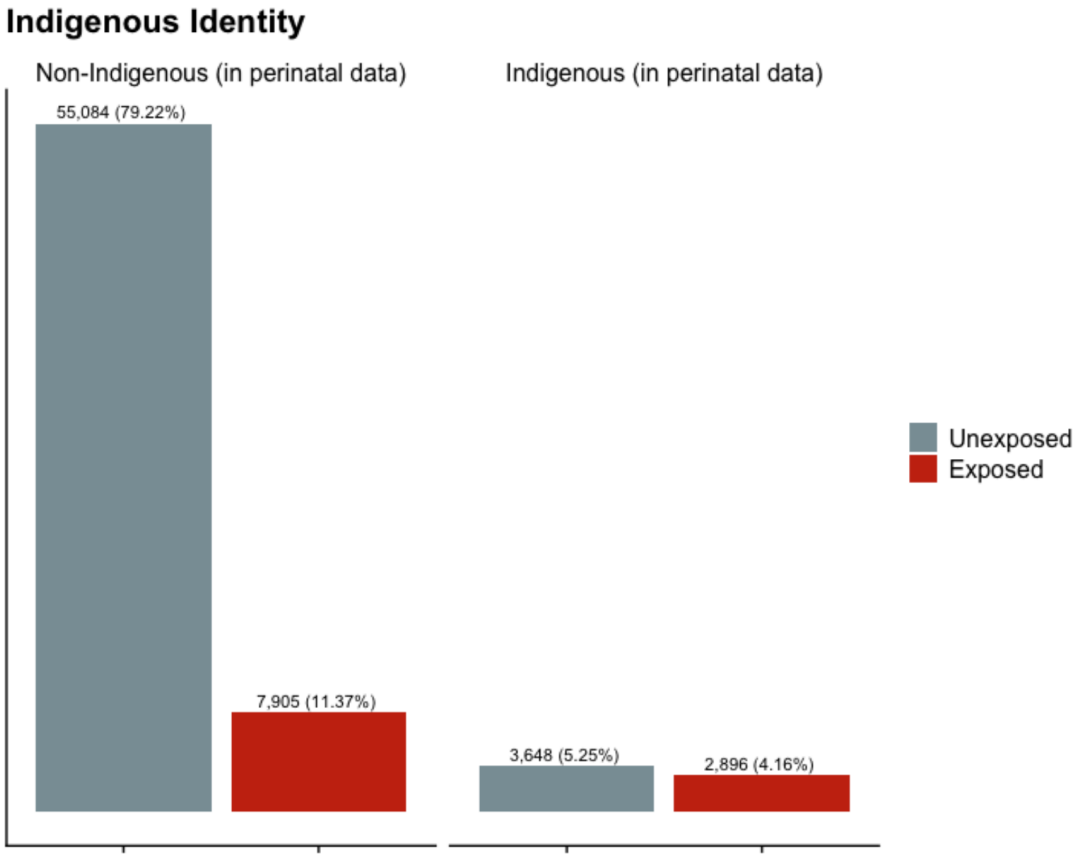

Figure 2.9: **Proportion of the cohort with a pregnancy in the outcome window that identified as Indigenous at pregnancy.**

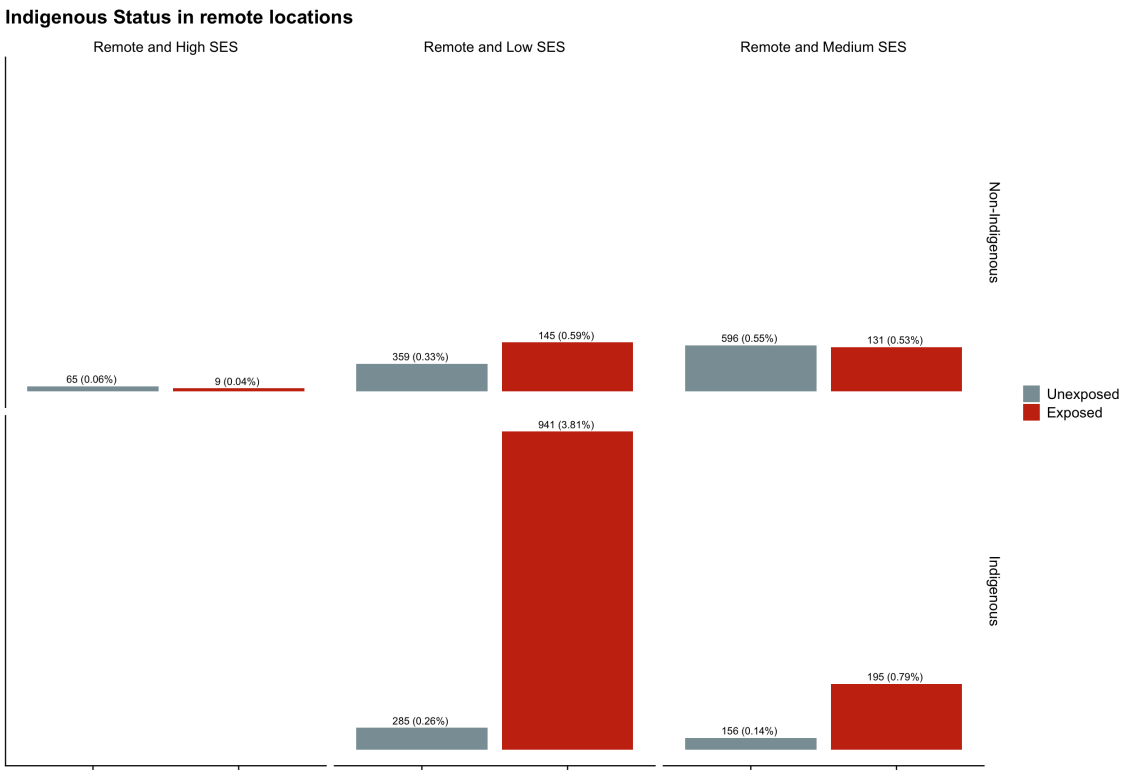

Figure 2.10: **Proportion of the cohort with a pregnancy in the outcome window that identified as Indigenous at pregnancy in remote locations.**

2.3.2 Preliminary Results and Model Comparisons

To fully account for the complexity of our dataset, we modelled the age effect with a restricted cubic spline, and included all two-way interaction effects between age, testing status and pregnancy in the testing window. In order to observe the influence of our modelling choices on our final results, we have also included the estimates and standard errors for a model with only main effects specified (Table 2.2), and a model with an interaction between testing history and prior pregnancy (Table 2.3).

Table 2.4 shows that the main effect for a having a history of testing, without a pregnancy in the testing period (as a measure of exposure to STIs), is 0.92 (95% CI 0.89 - 0.95) and that the estimate reduces once the interaction between women with a testing history and with a pregnancy in the testing window is accounted for (aOR 0.82; 95% CI 0.79 - 0.84). The main age effect estimated, without any interactions, is shown in Figure 2.11, and the location effects are summarised in Figure 2.12.

2.3.3 Pregnancy Outcomes

Our main analysis was to evaluate the probability of having a pregnancy in the outcome window, given a women's history of testing for Chlamydia and/or Gonorrhea in the testing period prior. All adjusted odds ratios presented are compared to the odds for women in the unexposed (never tested) group. Women in the exposed group with no prior pregnancy in the testing period (2000–2005) had varying odds of having a pregnancy depending on their age at the end of the outcome period (Figure 2.5), however in general women in the exposed group, in the middle of the age range for our cohort, had lower chances of pregnancy compared to women in the unexposed group, whereas the younger and older women in the cohort had similar pregnancy outcomes between exposed and unexposed groups. Younger women (aged 20 in 2005 with no pregnancy during testing window) had slightly reduced odds of pregnancy in the outcome window compared to same women in the unexposed group (aOR 0.91; 95% CI 0.87–0.95). For women aged 25 in 2005

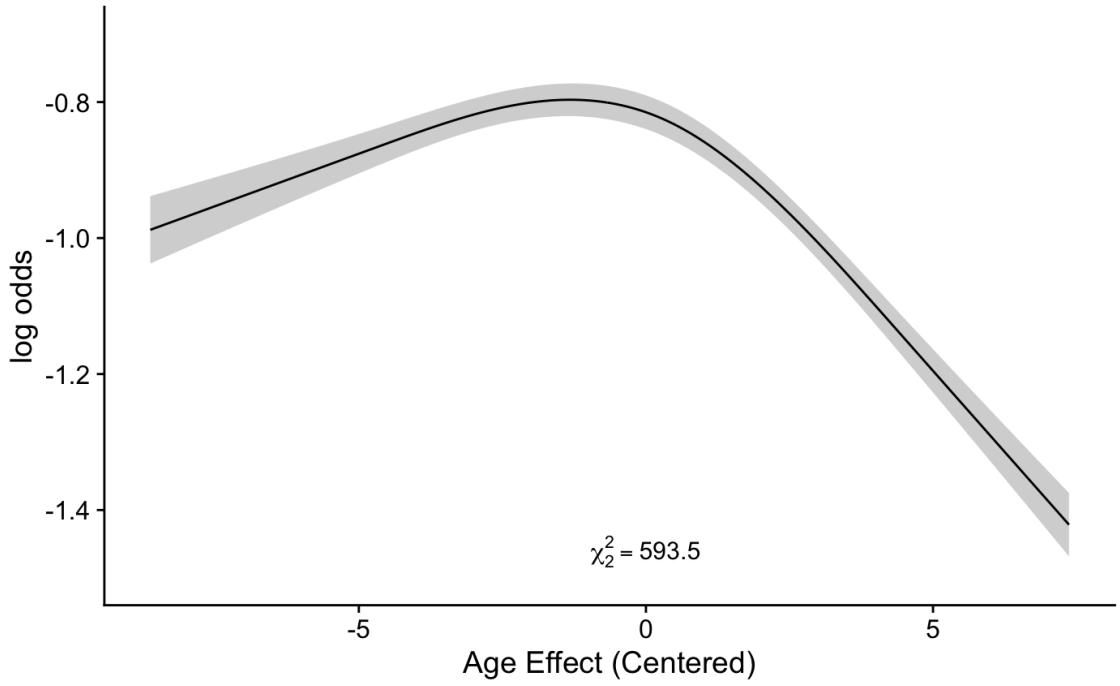

Figure 2.11: **Main age effect.** The age effect is estimated using a restricted cubic spline with 3 knots. The figure shows the age effect from the initial model with only main effects specified.

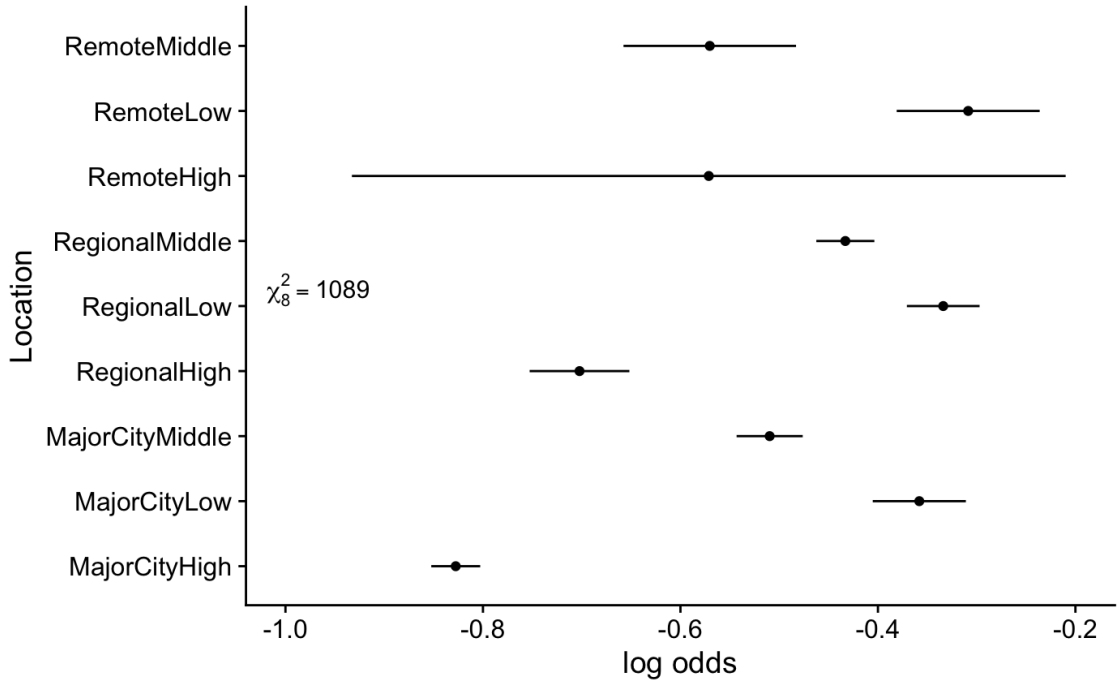

Figure 2.12: **Main location effects.** Location is defined as the interaction between the ABS defined regionality and socioeconomic status of the postcode of the initial testing event for the women in our dataset.

Variable	**Estimate**	**Standard Error**
Intercept	-0.7219	0.0158
History of testing	-0.0846	0.0157
Pregnant in the testing period	-0.1467	0.0140
Age (first basis function)	0.0308	0.0038
Age (second basis function)	-0.0758	0.0042
Location=MajorCityLow	0.4695	0.0250
Location=MajorCityMiddle	0.3179	0.0184
Location=RegionalHigh	0.1254	0.0267
Location=RegionalLow	0.4937	0.0197
Location=RegionalMiddle	0.3946	0.0163
Location=RemoteHigh	0.2563	0.1846
Location=RemoteLow	0.5190	0.0372
Location=RemoteMiddle	0.2573	0.0451

Table 2.2: **Summary of the model estimates, for the first model with only main effects specified.** Estimates are on the log-odds scale.

Variable	**Estimate**	**Standard Error**
Intercept	-0.7018	0.0158
History of testing	-0.2046	0.0175
Pregnant in the testing period	-0.2369	0.0151
History of testing and pregnant in the testing period	0.6090	0.0373
Age (first basis function)	0.0301	0.0038
Age (second basis function)	-0.0740	0.0043
Location=MajorCityLow	0.4749	0.0250
Location=MajorCityMiddle	0.3241	0.0184
Location=RegionalHigh	0.1277	0.0268
Location=RegionalLow	0.4939	0.0198
Location=RegionalMiddle	0.3922	0.0163
Location=RemoteHigh	0.2639	0.1847
Location=RemoteLow	0.5033	0.0372
Location=RemoteMiddle	0.2655	0.0451

Table 2.3: **Summary of the model estimates, for the second model with one interaction effect specified.** Estimates are on the log-odds scale.

Main Effects Model	
History of testing (exposed)	0.92 (0.89 - 0.95)
Unexposed	1 (ref)
Pregnant in the testing period	0.86 (0.84 - 0.89)
Not pregnant in the testing period	1 (ref)
Interaction Effect Model	
History of testing (exposed) and pregnant in the testing period	1.18 (1.08 - 1.29)
History of testing (exposed) and not pregnant in the testing period	0.82 (0.79 - 0.84)
Pregnant in the testing period and unexposed	0.79 (0.77 - 0.81)
Not pregnant in the testing period and unexposed	1 (ref)

Table 2.4: **Odds ratio for the testing and pregnancy effects, derived from the preliminary models and adjusted for age and location.**

the reduction in the chances of having a pregnancy was larger (aOR 0.71; 95% CI 0.68–0.75; Table 2.8). Women aged 30 years at 2005 (with no prior pregnancy in the testing window) had slightly reduced odds of pregnancy in the outcome window compared to women of the same age in the unexposed group (aOR 0.87; 95% CI 0.79–0.95).

When investigating the association between the classification of tests within the exposure group and pregnancy, we found that women with all negative tests had reduced odds of pregnancy compared to women in the unexposed group, but women with a single positive test had similar outcomes to the unexposed group, amongst women aged 20 in 2005 with no pregnancy in the testing period (all negative tests aOR 0.86; 95% CI 0.81–0.90, single positive test aOR 1.07; 95% CI 0.97–1.17). The relationship between prior pregnancy and pregnancy in the outcome window was similar for the older women in our cohort (30-year-old women and all negative tests aOR 0.87; 95% CI 0.78– 0.96, 30-year-old women and single positive test aOR 1.00; 95% CI 0.76–1.31). For women in the middle of the age range of our cohort, having a single positive test with no prior pregnancy resulted in the lowest

probability of pregnancy in the outcome window (Figure 2.14), and women in this age range with all negative tests had significantly reduced odds compared to women in the unexposed group (25-year-old women and all negative tests aOR 0.72; 95% CI 0.68–0.75, 25-year-old women and single positive test aOR 0.66; 95% CI 0.59–0.75). Women with multiple positive tests had probabilities of pregnancy in the outcome window that were increased or comparable with women in the unexposed group (20-year-old women aOR 1.51; 95% CI 1.20– 1.90, 25-year-old women aOR 1.04; 95% CI 0.81–1.35, 30-year-old women aOR 0.90; 95% CI 0.40–2.03). As the confidence intervals were wide for this subgroup, the age effect for women with multiple positive tests was not plotted but is summarise in Table 2.8. The association between the type of the infection for women with a positive test (chlamydia only, gonorrhea only, chlamydia and gonorrhea) showed that women aged 25 with no pregnancy in the testing period, had reduced odds of pregnancy in the outcome window if they tested positive for chlamydia (aOR 0.70; 95% CI 0.62–0.79). The odds ratio for women testing positive for gonorrhea in this group was higher and the confidence interval was wide (aOR 0.78; 95% CI 0.50–1.21). Women with positive results for both pathogens were comparable to all women in the exposed group, and the odds of pregnancy was influenced by age (Figure 2.15, Table 2.8).

Amongst women with a pregnancy in the testing period, women in the exposed group across all age ranges had increased odds of having a pregnancy in the outcome window, compared to women in the unexposed group. This difference was most pronounced for younger women (20-year-old women aOR 1.72; 95% CI 1.59–1.86, 25-year-old women aOR 1.35; 95% CI 1.26–1.45, 30-year-old women aOR 1.64; 95% CI 1.48–1.83). Women with a pregnancy in the testing period also had increased chances of a pregnancy in the outcome window if they had a single positive test compared to all negative tests (Figure 2.14, Table 2.8), and greater chances again if they had multiple positive tests. Women with a pregnancy during the testing window that had a gonococcal infection had increased chances of pregnancy in the outcome window compared to women with a chlamydial infection, although both groups of women with the two pathogens had reduced odds of pregnancy compared to women who were never tested.

Variable	Estimate	Standard Error
Intercept	-0.6613	0.0182
History of Testing	-0.4348	0.0347
Pregnant in the testing period	-0.2715	0.0309
Age (first basis function)	0.0495	0.0046
Age (second basis function)	-0.0866	0.0055
Location=MajorCityLow	0.4729	0.0250
Location=MajorCityMiddle	0.3216	0.0184
Location=RegionalHigh	0.1288	0.0268
Location=RegionalLow	0.4914	0.0198
Location=RegionalMiddle	0.3915	0.0163
Location=RemoteHigh	0.2570	0.1848
Location=RemoteLow	0.4959	0.0373
Location=RemoteMiddle	0.2604	0.0451
History of testing and pregnant in the testing period	0.6408	0.0384
History of testing and Age (first basis function)	-0.0827	0.0093
History of testing and Age (second basis function)	0.0798	0.0114
Pregnant in the testing period and Age (first basis function)	-0.0431	0.0115
Pregnant in the testing period and Age (second basis function)	0.0208	0.0111

Table 2.5: **Summary of the model estimates, for the full model with all interaction effects specified.** Estimates are on the log-odds scale.

Variable	Estimate	Standard Error
Intercept	-0.6596	0.0182
All Negative Tests	-0.4174	0.0373
Multiple Positive Tests	0.0513	0.2018
Single Positive Test	-0.6084	0.0877
Pregnant in the testing period	-0.2724	0.0309
Age (first basis function)	0.0495	0.0046
Age (second basis function)	-0.0868	0.0055
Location=MajorCityLow	0.4731	0.0250
Location=MajorCityMiddle	0.3223	0.0184
Location=RegionalHigh	0.1308	0.0268
Location=RegionalLow	0.4865	0.0198
Location=RegionalMiddle	0.3920	0.0163
Location=RemoteHigh	0.2589	0.1848
Location=RemoteLow	0.4423	0.0373
Location=RemoteMiddle	0.2524	0.0451
All negative tests and pregnant in the testing period	0.5798	0.0412
Multiple positive tests and pregnant in the testing period	1.0175	0.2136
Single positive test and pregnant in the testing period	0.9867	0.1040
All negative tests and Age (first basis function)	-0.0644	0.0102
All negative tests and Age (second basis function)	0.0674	0.0123
Multiple positive tests and Age (first basis function)	-0.0864	0.0535
Multiple positive tests and Age (second basis function)	0.0360	0.0787
Single positive test and Age (first basis function)	-0.1647	0.0212
Single positive test and Age (second basis function)	0.1619	0.0288
Pregnant in the testing period and Age (first basis function)	-0.0422	0.0116
Pregnant in the testing period and Age (second basis function)	0.0207	0.0111

Table 2.6: **Summary of the model estimates, for the full model with all interaction effects specified, where testing history is broken down by infection status.** Estimates are on the log-odds scale.

Variable	Estimate	Standard Error
Intercept	-1.1959	0.0943
Positive for both pathogens (CT/NG)	0.2128	0.2173
All negative tests	0.1270	0.1001
Never tested	0.5506	0.0951
Positive for gonorrhea only	0.2953	0.3512
Pregnant in the testing window	0.5689	0.1087
Age (first basis function)	-0.0993	0.0229
Age (second basis function)	0.0627	0.0337
Location=MajorCityLow	0.4731	0.0250
Location=MajorCityMiddle	0.3228	0.0184
Location=RegionalHigh	0.1310	0.0268
Location=RegionalLow	0.4887	0.0198
Location=RegionalMiddle	0.3918	0.0163
Location=RemoteHigh	0.2601	0.1848
Location=RemoteLow	0.4420	0.0382
Location=RemoteMiddle	0.2529	0.0452
Both pathogens (CT/NG) and pregnant in the testing period	0.6611	0.2508
All negative tests and pregnant in the testing period	-0.2702	0.1116
Never tested and pregnant in the testing period	-0.8499	0.1060
Positive for gonorrhea only and pregnant in the testing period	0.9069	0.4749
Both pathogens (CT/NG) and Age (first basis function)	-0.0028	0.0499
All negative tests and Age (first basis function)	0.0868	0.0247
Never tested and Age (first basis function)	0.1516	0.0233
Positive for gonorrhea only and Age (first basis function)	0.0749	0.0895
Both pathogens (CT/NG) and Age (second basis function)	-0.0332	0.0829
All negative tests and Age (second basis function)	-0.0870	0.0359
Never tested and Age (second basis function)	-0.1598	0.0341
Positive for gonorrhea only and Age (second basis function)	-0.1232	0.1168
Pregnant in the testing period and Age (first basis function)	-0.0443	0.0118
Pregnant in the testing period and Age (second basis function)	0.0248	0.0123

Table 2.7: **Summary of the model estimates, for the full model with all interaction effects specified, where testing history is broken down by infection type.** Estimates are on the log-odds scale.

		aOR[a] (95% CI)		aOR (95% CI)		aOR (95% CI)
Not Pregnant in the testing period						
			Multiple Positive Tests (n = 468)	1.51 (1.20 , 1.90)	CT (n = 2,500)	1.08 (0.98, 1.20)
Aged 20 at end of testing period	Exposed[b] (n = 20, 250)	0.91 (0.87, 0.95)	Single Positive Test (n = 2,756)	1.07 (0.97, 1.17)	NG (n = 171)	1.06 (0.71, 1.59)
			All Negative Tests (n = 17,026)	0.86 (0.82, 0.90)	CT/NG (n = 553)	1.35 (1.11, 1.65)
			Multiple Positive Tests	1.04 (0.81, 1.35)	CT	0.70 (0.62, 0.79)
Aged 25		0.71 (0.68, 0.75)	Single Positive Test	0.66 (0.59, 0.75)	NG	0.78 (0.50, 1.21)
			All Negative Tests	0.72 (0.68, 0.75)	CT/NG	0.81 (0.63, 1.04)
			Multiple Positive Tests	0.90 (0.40, 2.03)	CT	1.01 (0.76, 1.36)
Aged 30		0.87 (0.79, 0.95)	Single Positive Test	1.00 (0.76, 1.31)	NG	0.69 (0.30, 1.60)
			All Negative Tests	0.87 (0.78, 0.96)	CT/NG	0.91 (0.43, 1.93)
	Unexposed (n = 75, 217)	1 (ref)	Unexposed	1 (ref)	Unexposed	1 (ref)
Pregnant in the testing period						
			Multiple Positive Tests (n = 126)	4.17 (2.72, 6.40)	CT (n = 463)	2.53 (2.06, 3.11)
Aged 20 at end of testing period	Exposed (n = 4,452)	1.72 (1.59, 1.86)	Single Positive Test (n = 481)	2.85 (2.33, 3.40)	NG (n = 27)	6.17 (2.47, 15.40)
			All Negative Tests (n = 3,845)	1.53 (1.40, 1.67)	CT/NG (n = 117)	6.13 (3.92, 9.59)
			Multiple Positive Tests	2.89 (1.97, 4.24)	CT	1.64 (1.35, 1.99)
Aged 25		1.35 (1.26, 1.45)	Single Positive Test	1.77 (1.47, 2.15)	NG	4.53 (1.84, 11.14)
			All Negative Tests	1.28 (1.19, 1.38)	CT/NG	3.66 (2.43, 5.51)
			Multiple Positive Tests	2.50 (1.07, 5.83)	CT	2.37 (1.70, 3.31)
Aged 30		1.64 (1.48, 1.83)	Single Positive Test	2.67 (1.95, 3.65)	NG	3.99 (1.38, 11.60)
			All Negative Tests	1.55 (1.39, 1.74)	CT/NG	4.14 (1.88, 9.13)
	Unexposed (n = 33,043)	1 (ref)	Unexposed	1 (ref)	Unexposed	1 (ref)

Table 2.8: **Odds ratios of pregnancy in the outcome period.**

[a] All models are adjusted for location and pregnancy in the testing period interacted with a cubic spline of age evaluated at 20, 25 and 30 (as at 2005).

[b] Any history of testing for chlamydia/gonorrhea.

To further evaluate the reproductive burden associated with chlamydia and gonorrhea, we replicated our pregnancy analysis using perinatal outcomes of low birth-weight and pre-term birth. The women who had a history of chlamydia and/or gonorrhea testing were more likely to have adverse perinatal outcomes such as low birth-weight and pre-term birth (Table 2.11; low birth-weight aOR 1.35; 95% CI 1.23–1.48, pre-term birth aOR 1.28; 95% CI 1.18–1.39). Women with multiple positive tests had significantly higher odds of the low birth-weight outcome compared to women with single positive tests and all negative tests (multiple positive tests aOR 2.61; 95% CI 1.85–3.69, single positive test 1.50; 95% CI 1.23–1.84, all negative tests aOR 1.29; 95% CI 1.17– 1.43). The odds of pre-term birth were comparable between all testing groups.

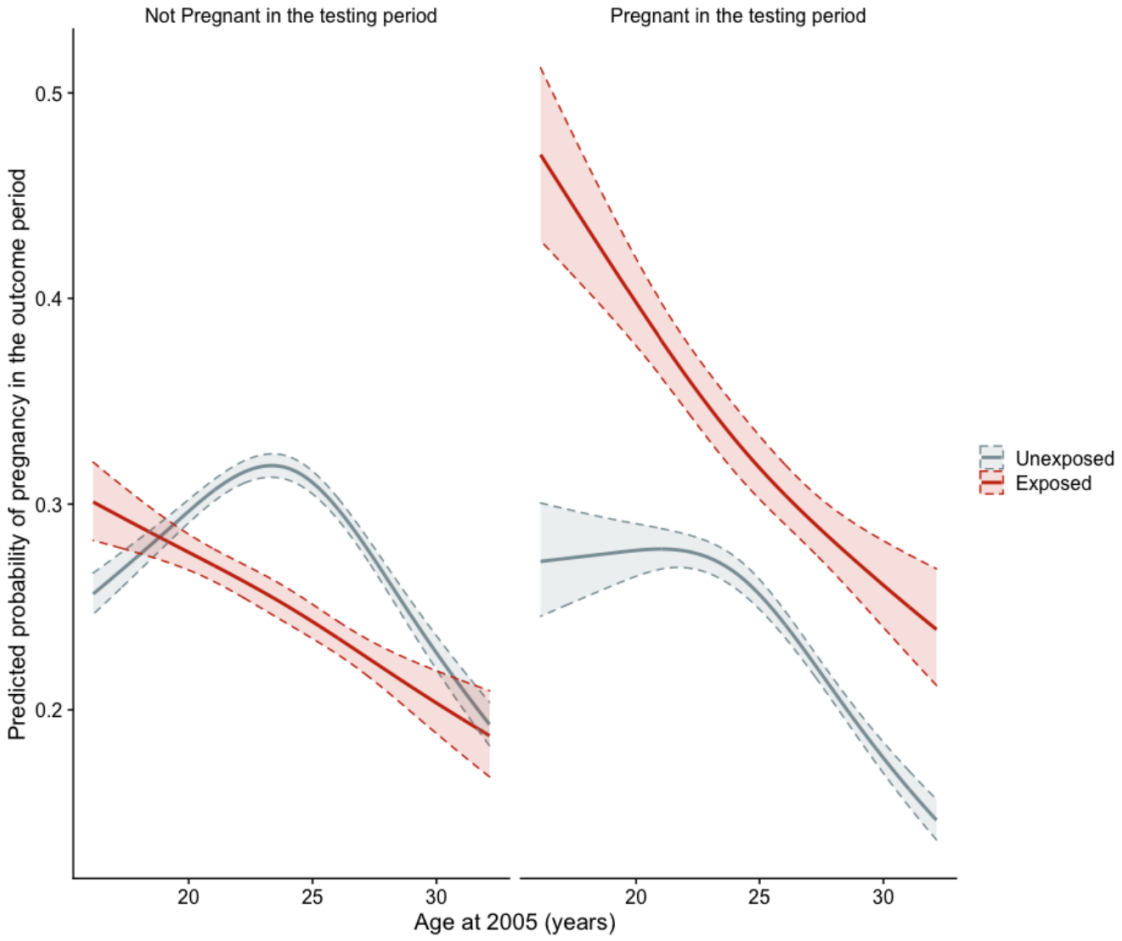

Figure 2.13: **Association of chlamydia and gonorrhea testing and pregnancy.** Compares exposed and unexposed group. All lines represent the predicted probability of having a pregnancy in the outcome period, and the shaded area represents the (endpoint transformed) 95% confidence interval. Model adjusted to major city and high SES locations.

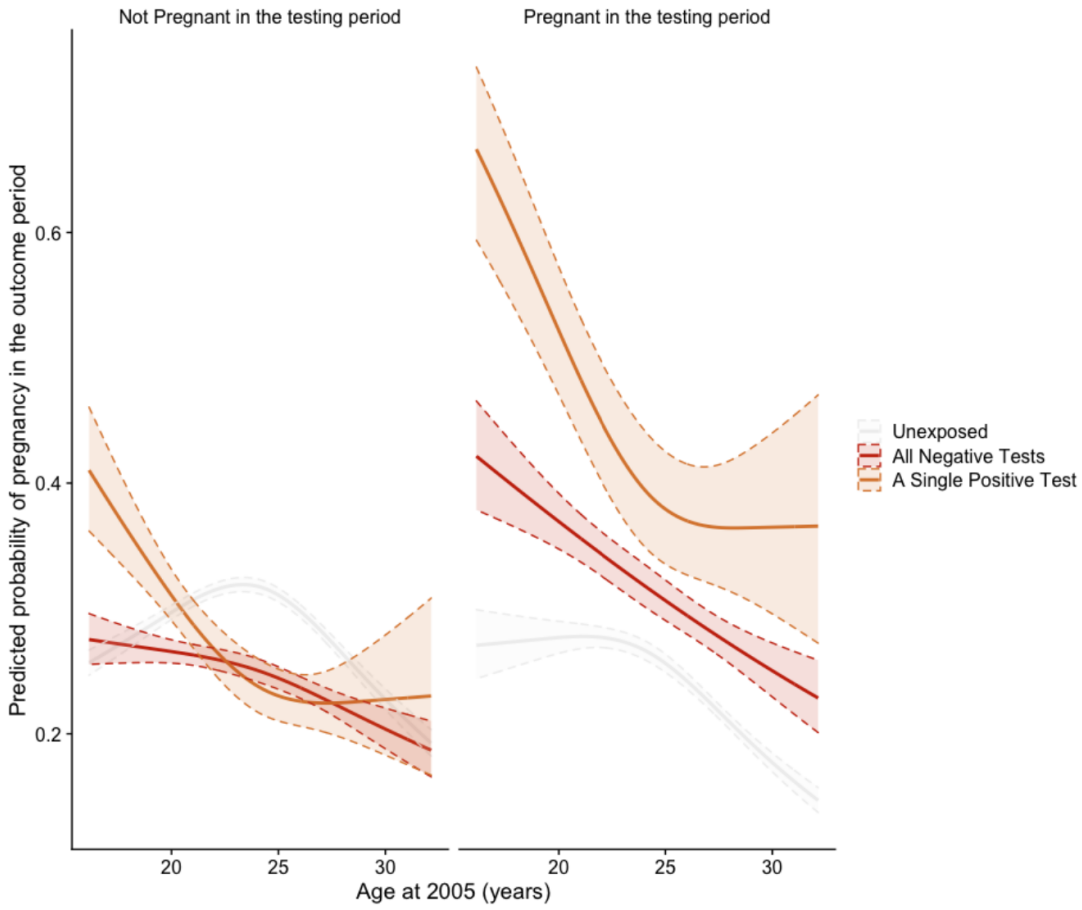

Figure 2.14: **Association of chlamydia and gonorrhea testing and pregnancy.** (Compares single positive test, all negative tests and unexposed group. All lines represent the predicted probability of having a pregnancy in the outcome period, and the shaded area represents the (endpoint transformed) 95% confidence interval. Model adjusted to major city and high SES locations.

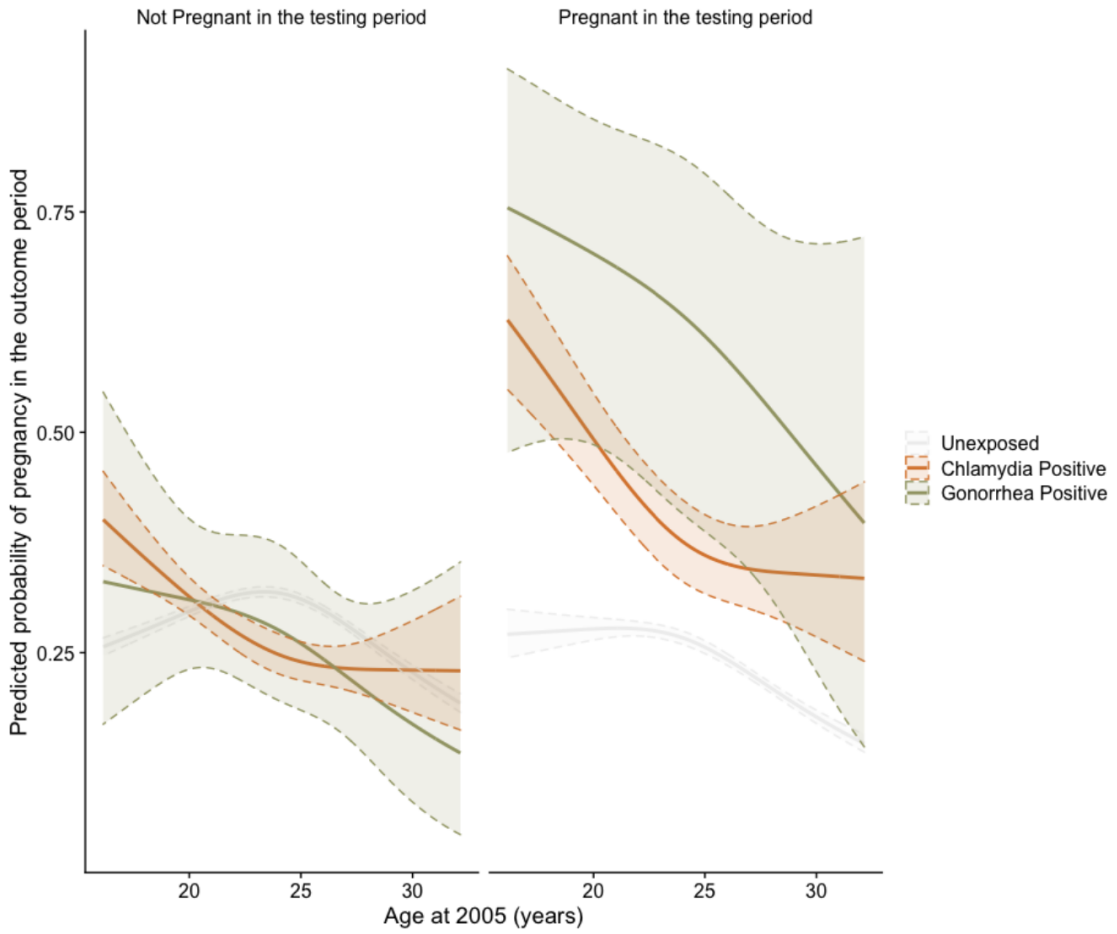

Figure 2.15: **Association of chlamydia and gonorrhea testing and pregnancy.** (Compares Chlamydia positive, Gonorrhea positive and unexposed group. All lines represent the predicted probability of having a pregnancy in the outcome period, and the shaded area represents the (endpoint transformed) 95% confidence interval. Model adjusted to major city and high SES locations.

2.3.4 Hierarchical pregnancy models

It is clear that a history of testing is associated with reduced odds of pregnancy in the outcome window. Testing history is associated with higher risk of infection (for the subgroup with negative tests; for those with positive tests an infection history is also present), and testing history is associated with being from lower SES areas, which may also independently impact pregnancy outcomes (Figure 2.12). Testing history may also be correlated with other biological factors that are not explicitly represented in our data. To evaluate the extent to which the reproductive burden of testing history is an indication of demographic or underlying biological factors, we fit a series of hierarchical (or multilevel) models with postcode specific terms to adjust for demographic factors associated with a particular area. We fit three models, a fixed effects model with no postcode-level intercept (this replicates the main model discussed in the previous section), a varying intercept model with a postcode-level intercept, and a varying effects model with a postcode-level intercept and with the effect of testing history and prior pregnancy allowed to vary across postcodes. Models are fit within a Bayesian framework, as using Markov Chain Monte Carlo estimation ensures convergence of the models. Instead of restricted cubic splines, thin plate splines are used in the models. For the coefficient parameters of the models, the standard deviation of the smooth terms and the standard deviation of the varying intercepts and effects are given N(0,1) priors. Figure 2.16 replicates the results shown in Figure 2.13. There are some differences between plots due to the greater degrees of freedom allowed to the spline terms in the later model. The overall difference between testing and unexposed groups have been described in previous sections, and the results of this model is similar. For example, the adjusted odds ratio for women with a testing history and no prior pregnancy is 0.89 (0.87 - 0.91), 0.71 (0.70 - 0.73) and 0.85 (0.80 - 0.89) for women aged 20, 25 and 30 years respectively. Once area-level variability is adjusted for using postcode-specific intercepts in a varying intercept hierarchical model, the odds ratio for the exposed compared to the unexposed group, for women with no prior pregnancy, shifts slightly towards unity across all ages (Table **??**, Figure 2.17). For example, for women aged 20 at 2005, the odds ratio was 0.94 (0.93 - 0.96) in the

varying intercept model, where it was 0.89 (0.87 - 0.91) for the fixed effects model, and similarly for women across all ages. The odds ratio for women with a prior pregnancy shifts away from unity, becoming larger. For women aged 25, the varying intercept odds ratio was 1.45 (1.42 - 1.48), whereas it was 1.35 (1.30 - 1.39) in the fixed effects model. The inclusion of this effect, compared to the one-level fixed effects model, suggests that a small part of the testing history effect observed in the previous section is attributable to postcode related demographic factors, but a majority of the effect is explained by alternative, biological factors. The distribution of postcode-level intercepts is shown in Figure 2.20. The standard deviation of the postcode-level parameters is 0.37, this would imply that the odds of pregnancy is 1.45 for a woman of average age, that has never been tested and with no prior pregnancy in a postcode that has an intercept of one standard deviation above the average postcode, and 0.69 of a woman in a postcode that is one standard deviation below the average. After allowing the effect of testing history (exposure) and prior pregnancy to vary across postcode, we observe the difference between exposed and unexposed groups, for women with no prior pregnancy, becomes smaller for middle age ranges and close to unity for age ranges on the extremes (Figure 2.18, Table 2.9). The average odds ratio, for women with a prior pregnancy, is smaller than the effects observed from the varying intercept model. The comparisons shown are for a predictions drawn from the distributions of effects across all postcodes. There are three interesting correlations of the varying effects that can be found in Table 2.10. The correlation between the Intercept and the testing history effect between postcodes is -0.16 (95% CrI -0.31, 0.01) which suggests that postcode with overall increased pregnancy rates will have larger negative effects associated with exposure. Similarly, the correlation between the Intercept and Prior Pregnancy is -0.85 (95% CrI -0.90, -0.79). The correlation between the exposure effect and the interaction between exposure and prior pregnancy is -0.75 (-0.86, -0.61). Figure 2.21 shows the distribution of total random effects (history of testing, prior pregnancy and intercept) across all postcodes. A majority of the effects cover zero on the log-odds scale, suggesting that the postcode effects for these areas are negligible. Some of the postcodes has effects that are significantly less than zero (on the log-odds scale), implying that these are postcodes with factors that significantly lower overall chances

		Fixed Effects Model aOR (95% CrI)	Varying Intercept Model aOR (95% CrI)	Varying Effects Model aOR (95% CrI)
Age 20 years (2005)	No prior pregnancy	0.89 (0.87 - 0.91)	0.94 (0.93 - 0.96)	1.06 (1.04 - 1.09)
	Pregnant in the testing period	1.69 (1.65 - 1.74)	1.83 (1.80 - 1.87)	1.77 (1.69 - 1.85)
Age 25 years (2005)	No prior pregnancy	0.71 (0.70 - 0.73)	0.75 (0.74 - 0.76)	0.80 (0.78 - 0.82)
	Pregnant in the testing period	1.35 (1.30 - 1.39)	1.45 (1.42 - 1.48)	1.33 (1.26 - 1.40)
Age 30 years (2005)	No prior pregnancy	0.85 (0.80 - 0.89)	0.87 (0.84 - 0.90)	0.87 (0.83 - 0.92)
	Pregnant in the testing period	1.61 (1.51 - 1.71)	1.69 (1.61 - 1.77)	1.46 (1.36 - 1.58)

centering

Table 2.9: **Odds ratios of pregnancy in the outcome period.** All models are adjusted for location and pregnancy in the testing period interacted with a spline of age evaluated at 20, 25 and 30 (as at 2005).

of pregnancy. Figure 2.19 shows the age effect for exposed and unexposed groups by postcode. Using the estimates in Table 2.10, we can compute the standard deviation of the varying effects for all four groups of history of testing and prior pregnancy. For woman with no history of testing and no prior pregnancy, the standard deviation of the postcode-level effects is 0.49, implying that a woman of average age in a postcode 1 standard deviation below the mean would have an odds ratio of pregnancy of 0.63, and a woman in a postcode 1 standard deviation above the average would have odds of pregnancy 1.60. For a woman with a history of testing and no prior pregnancy, the standard deviation is 0.67 (95% CrI 0.58 - 0.78), and so odds of pregnancy in postcodes 1 standard deviation below and above the average would be 0.51 and 1.95. For women with no history of testing and with prior pregnancy, the standard deviation is 0.62 (95% CrI 0.55 - 0.69) and so odds of pregnancy 1 standard deviation below and above the mean would be 0.54 and 1.85. For women with a history of testing and with prior pregnancy, the standard deviation is 0.93 (95% CrI 0.82 - 1.06), and so women in postcodes 1 standard deviation below and above the mean would have odds of pregnancy 0.39 and 2.53. The variability related to postcode-specific demographic factors is smallest for women with no history of testing and no prior pregnancy, and higher for women with a history of testing and women with a prior pregnancy. The variability is largest for women with both a history of testing and prior pregnancy, suggesting that the effect of exposure is highly positive within some postcodes, and highly negative in others. The postcode-level variation also outweighs most of the population level effects found in the model.

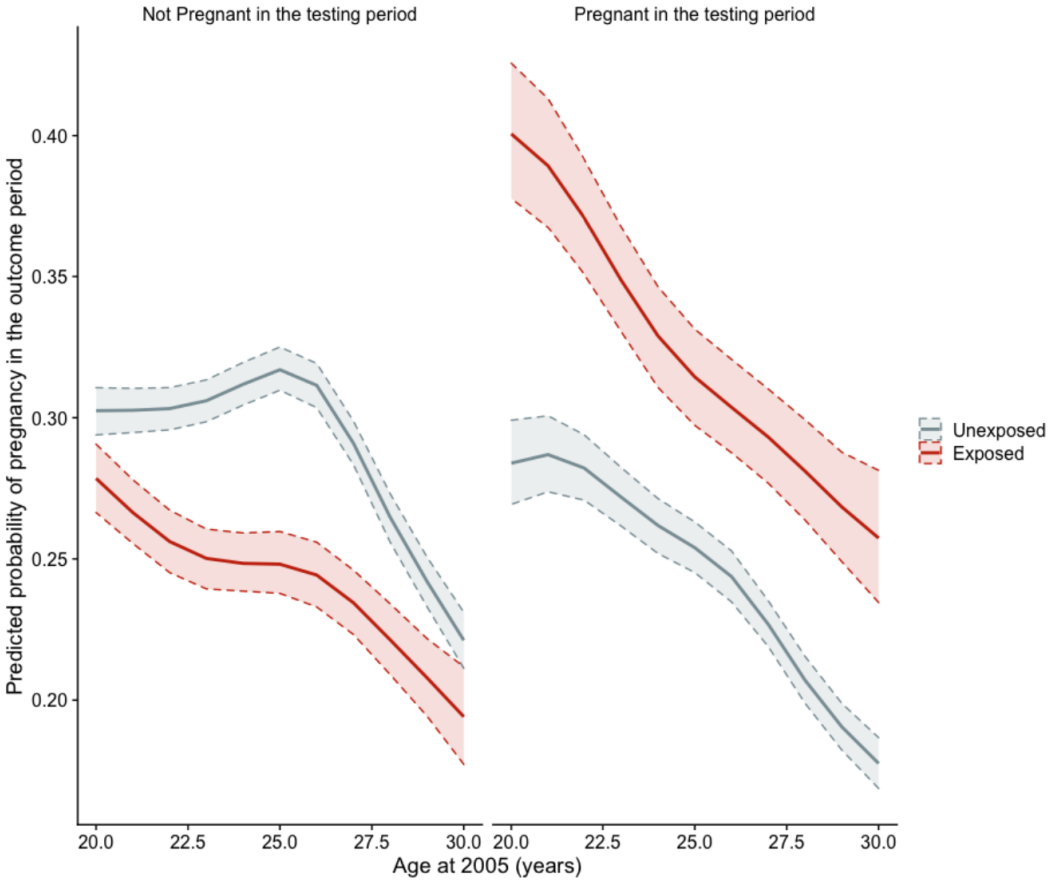

Figure 2.16: **Association of chlamydia and gonorrhea testing and pregnancy.** Compares exposed and unexposed group. All lines represent the predicted probability of having a pregnancy in the outcome period, and the shaded area represents the 95% posterior interval. Adjusted to major city and high SES locations.

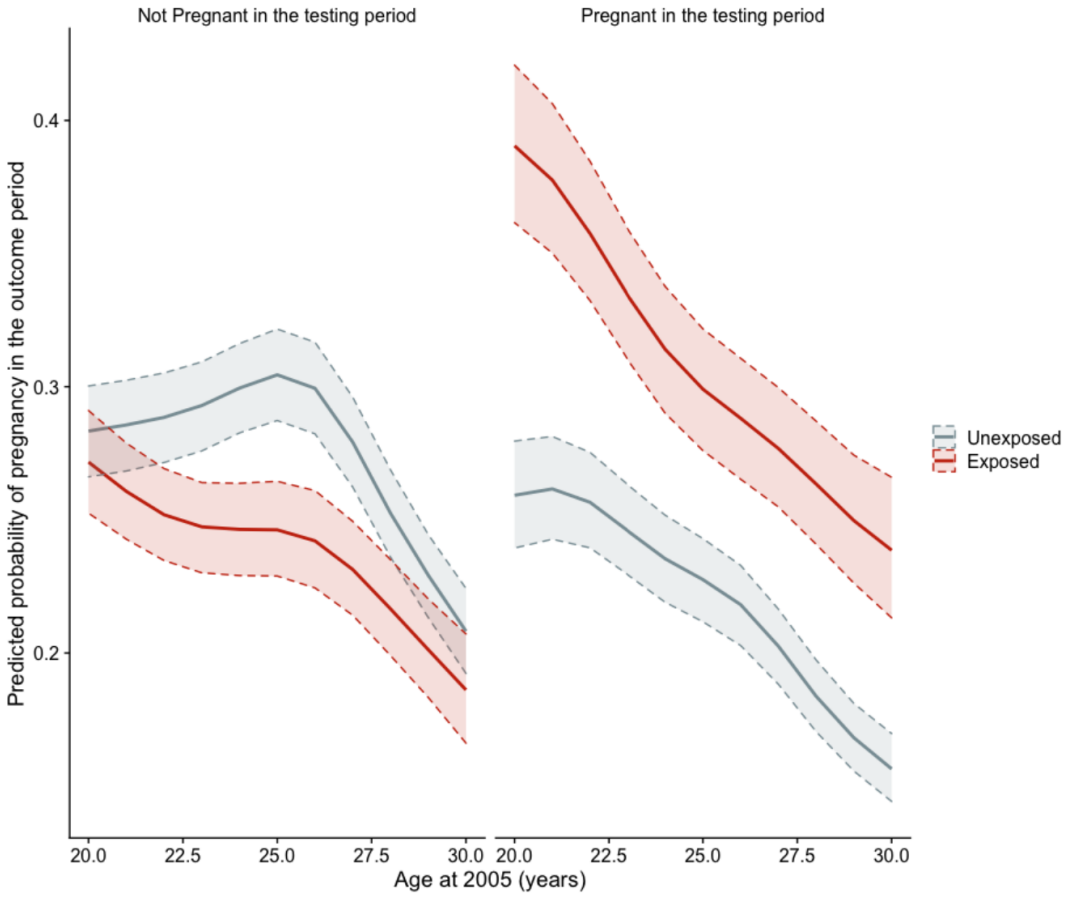

Figure 2.17: **Association of chlamydia and gonorrhea testing and pregnancy.** Compares exposed and unexposed group. Model is adjusted for postcode level effects using a varying intercept term. All lines represent the predicted probability of having a pregnancy in the outcome period, and the shaded area represents the 95% posterior interval. Adjusted to major city and high SES locations.

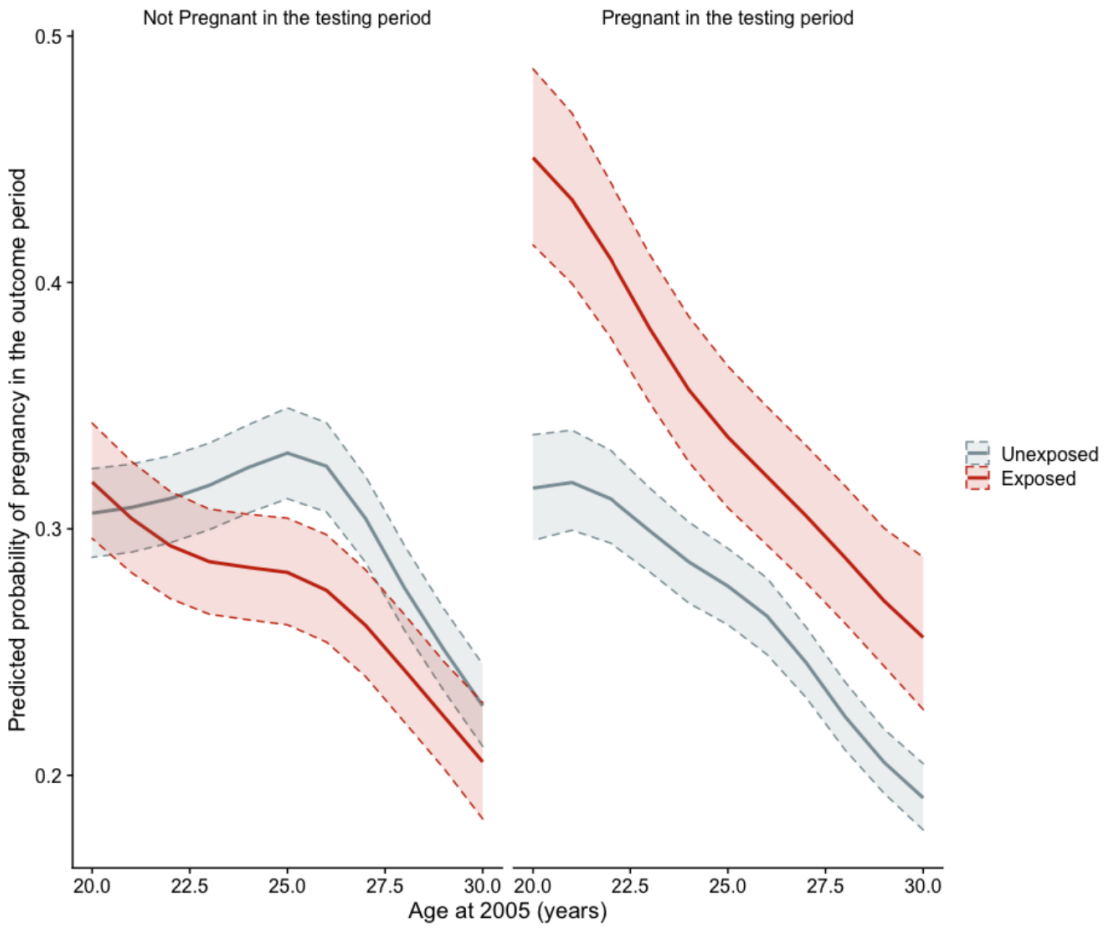

Figure 2.18: **Association of chlamydia and gonorrhea testing and pregnancy.** Compares exposed and unexposed group. Model is adjusted for postcode level effects using a varying intercept term, and the effects of testing and prior pregnancy is allowed to vary across postcodes. All lines represent the predicted probability of having a pregnancy in the outcome period, and the shaded area represents the 95% posterior interval. Adjusted to major city and high SES locations.

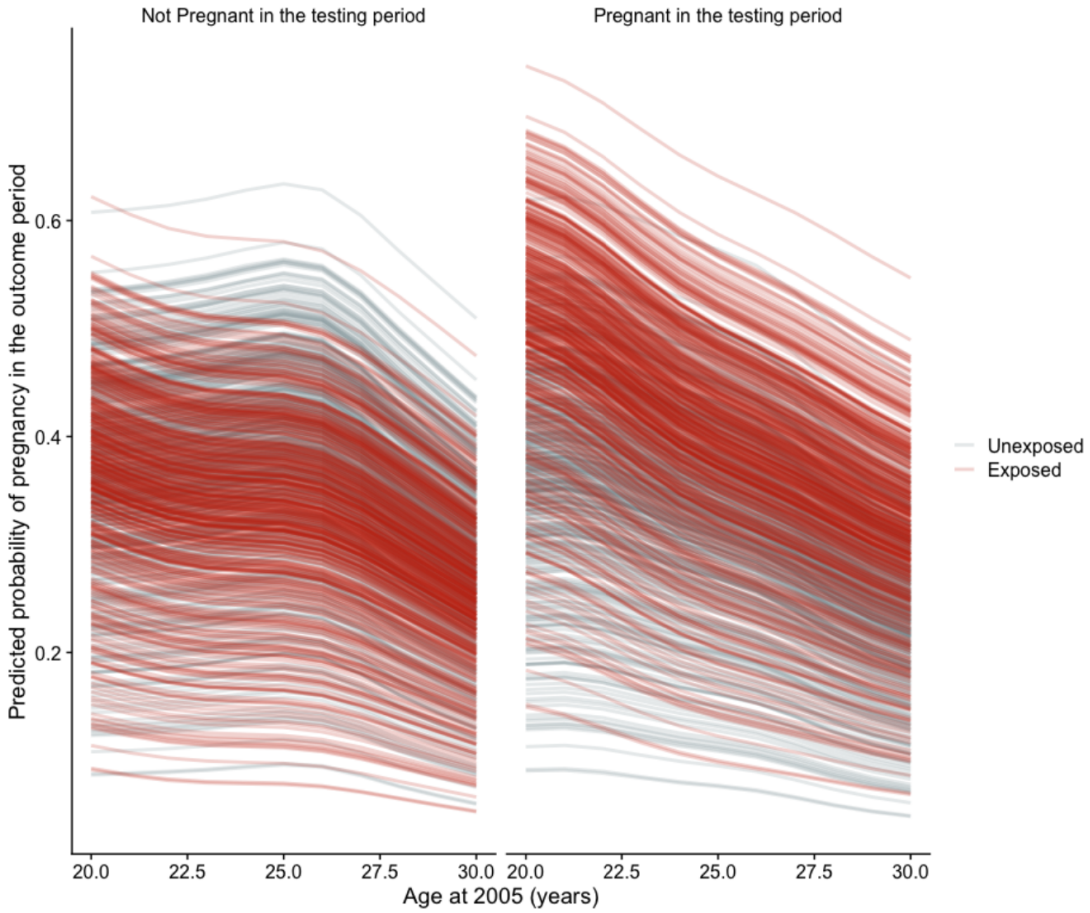

Figure 2.19: **Association of chlamydia and gonorrhea testing and pregnancy.** Compares exposed and unexposed group. Model is adjusted for postcode level effects using a varying intercept term, and the effects of testing and prior pregnancy is allowed to vary across postcodes.. Each line represents the predicted probability of having a pregnancy in the outcome period for a particular postcode. Adjusted to major city and high SES locations.

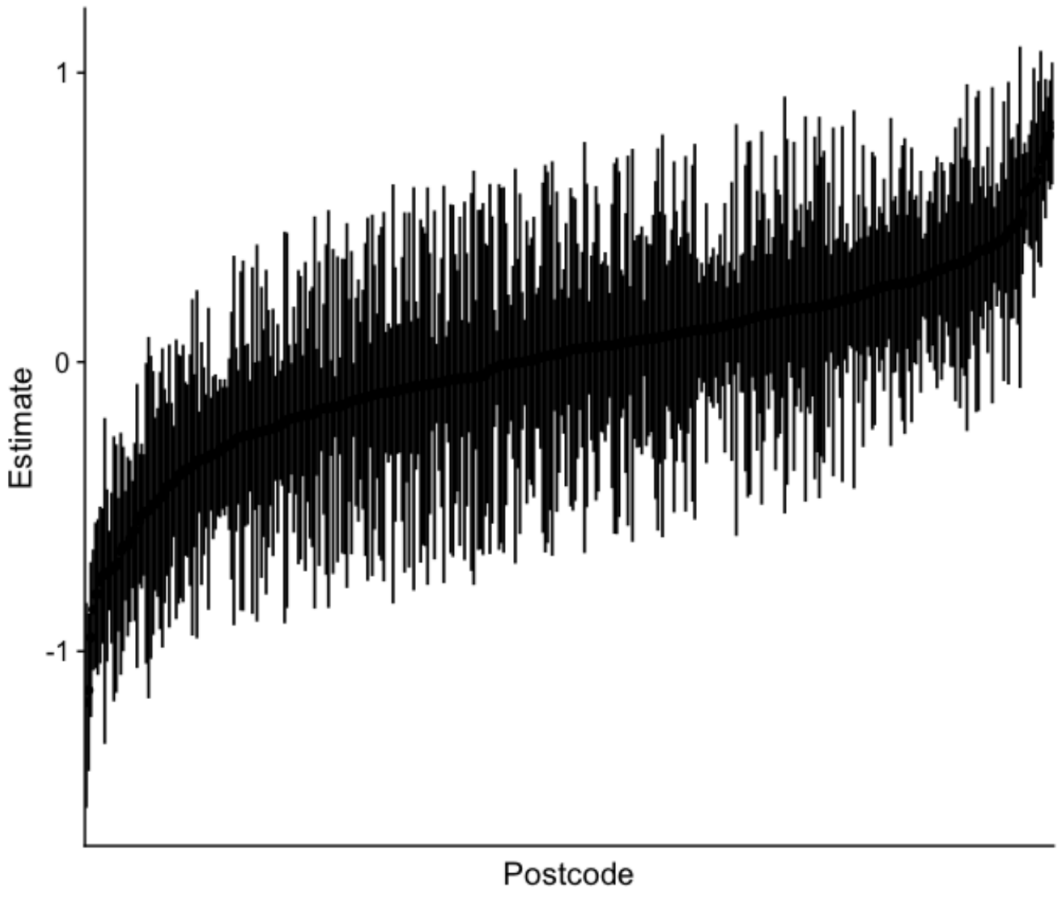

Figure 2.20: **Distribution of the postcode-level intercepts.** Points represent the mean of the posterior distribution for each random effect, and lines represent the 95% posterior interval.

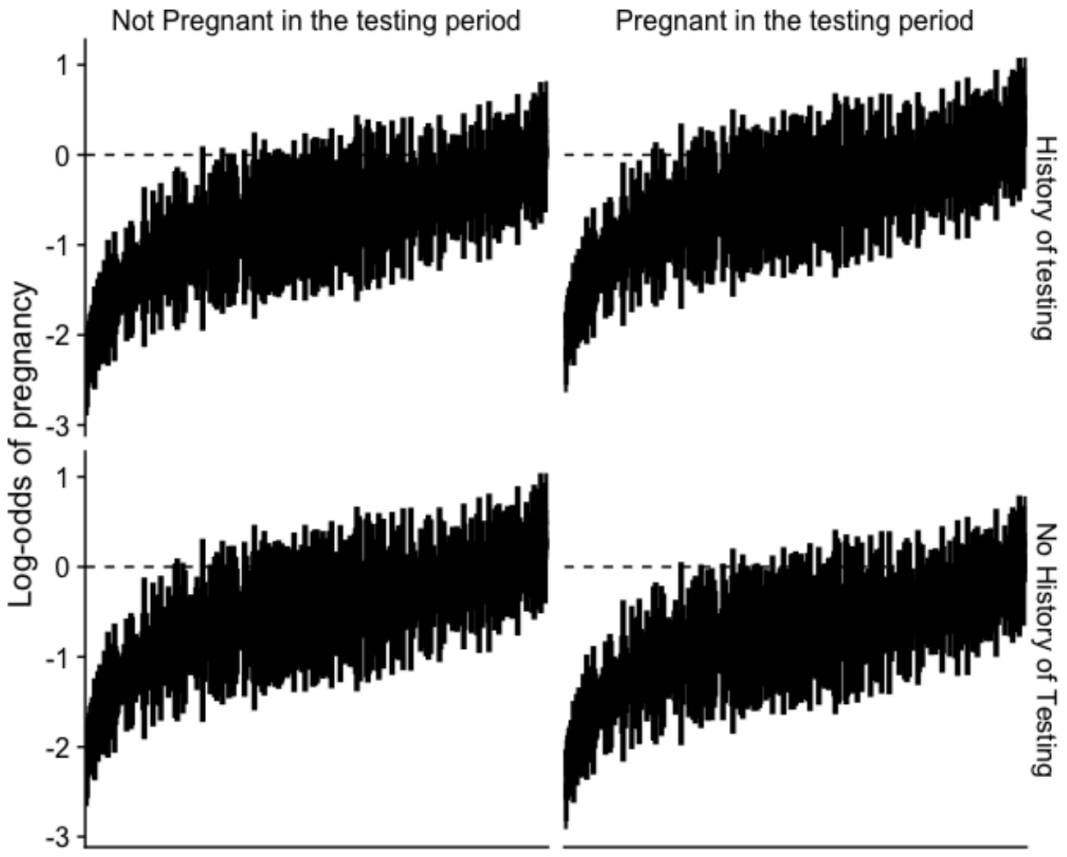

Figure 2.21: **Distribution of the postcode-level effects across key varying covariates**. Points represent the mean of the posterior distribution for each random effect, and lines represent the 95% posterior interval.

sd(Intercept)	0.49 (0.44, 0.53)
sd(History of testing)	0.48 (0.42, 0.54)
sd(Prior Pregnancy)	0.51 (0.43, 0.57)
sd(History of Testing and Prior Pregnancy)	0.60 (0.48, 0.72)
cor(Intercept, History of Testing)	-0.16 (-0.32, 0.01)
cor(Intercept, Prior Pregnancy)	-0.85 (-0.90, -0.79)
cor(History of Testing, Prior Pregnancy)	0.03 (-0.15, 0.20)
cor(Intercept, History of Testing and Prior Pregnancy)	-0.06 (-0.27, 0.16)
cor(History of Testing, History of Testing and Prior Pregnancy)	-0.75 (-0.86, -0.61)
cor(Prior Pregnancy, History of Testing and Prior Pregnancy)	0.17 (-0.07, 0.40)

Table 2.10: **Parameters of the postcode-level varying effects.** Estimates shown are the mean of the posterior distribution, and the 95% posterior interval for each parameter is displayed in brackets.

2.3.5 Discrete-time pregnancy models

It appears that the age effects are somehow accounting for the effects of time from first test until first pregnancy, which is why it varies so much between testing groups. To investigate this, we run a discrete time analysis on the time until first pregnancy. Because we lack a suitable starting point for the whole cohort, we model women in the 15-19 age range. Being in this age range at the end of the testing period still represents a reasonable starting point for time to first pregnancy. We exclude anyone that had a pregnancy prior to this date. We model the chances of a pregnancy within each year of follow up. As pregnancy records are reliable from 2007 (QLD used electronically collected perinatal records from the beginning of this year), we use this year as our first time period for modelling. We stop the follow up in 2012, as this was our last full year of follow up in the study. Women that have had a pregnancy are excluded from later time periods. Our first fit uses testing history as our exposure definition, with no interaction with time. We adjust the model for Location, and use no intercept term (so that the first period is the reference case for estimates). The models in this section are fitted using the Bayesian framework, as this easily allows for the inclusion of uncertainty intervals on the survival curves, and is easy to include person-specific intercepts (that is, easily extensible to a frailty

model). The parameters of the model are given N(0,1) priors. The estimate for the exposure group is negative (Figure 2.20). The hazard-odds ratio is 0.93 (95% CrI 0.90 - 0.97), which is assumed to be constant throughout the observation period. The survival curve is almost linear (Figure 2.21). We next model time until pregnancy, with an interaction between exposure and time. The hazard-odds ratio for the exposed group is 0.84 (95% CrI 0.76 - 0.92) - however this now applies to the first time period (2007). The hazard odds-ratio is less than unity for all years except for 2012, where the exposed group has greater probability of pregnancy. However, in most of those years the confidence intervals for the hazards significantly overlap (Figure 2.22). This effect is clearer in the survival curve (Figure 2.23), where it shows the exposed group 'catching up' to the unexposed group. This is consistent with the idea that exposure to testing reduces the ease at which women become pregnant and therefore increases the time to pregnancy. As discussed in 2.3.4, the exposure effect of testing varies significantly across postcodes. We fit the previous model, and include a postcode level intercept, and allow the effect testing history to vary across postcodes. The hazard-odds ratio for the exposed group, within the average postcode at 2007 is identical to the previous model - 0.84 (95% CrI 0.76 - 0.92). As in the previous model, the mean hazard-odds ratio is less than unity for all years except for 2012, where the exposed group has greater probability of pregnancy. The standard deviation of the postcode-level intercepts is 0.05 (95% CrI 0.01 - 0.08), so the hazard-odds ratio for a woman with no history of testing in a postcode 1 standard deviation below the mean would be 0.95, and in a postcode 1 standard deviation above the mean the hazard-odds ratio would be 1.05. The standard deviation of the postcode effects for women with a history of testing is 0.09 (95% CrI 0.00 - 0.24), which implies that the hazard-odds ratio for a woman in a postcode 1 standard deviation below the mean is 0.92, and in a postcode 1 standard deviation above the mean is 1.08. As observed previously with the hierarchical models fit to all pregnancies, the postcode level variability is greater for women with a history of testing, than for those with no history.

We can then finally model time until pregnancy, with the exposed group broken into testing diagnosis. Having a single positive test results in lower hazard-odds than having all negative tests (for single positive tests the hazard-odds ratio

is 0.73, 95% CrI 0.58 - 0.92 and for all negative tests, the hazard-odds ratio is 0.86, 95% CrI 0.78 - 0.95). There is a modest interaction between having all negative tests and time, which shows an increase in the chances of pregnancy as time goes on (Figure 2.29). The bounds for the multiple positive test effect is particularly large, and so no difference in pregnancy outcomes could be inferred for this group. Overall, a breakdown of time to pregnancy results by testing diagnosis showed a similar effect to the results presented earlier. Chapter 3 will discuss in greater detail the difference between the negative and positive testing groups.

Whilst our data used in this chapter has significant limitations that prevents us from doing a full discrete-time based analysis in a longitudinal framework, these results demonstrate the potential importance of considering the time-based effects. There is great future opportunity in gathering full longitudinal used to investigate more precise hypotheses about the relationship between testing, pathogens and reproductive burden. For example, as seen above the impact of testing may be more apparent in younger women or closer to the testing date.

2.3.6 Perinatal Outcomes

To further evaluate the reproductive burden associated with chlamydia and gonorrhea, we replicated our pregnancy analysis using perinatal outcomes of low birth-weight and pre-term birth. We present both outcomes simultanenously, as there was a large correlation between the two outcomes (birth weight and gestational weeks had a correlation of 0.67, 72.9% of low birth weight births were also pre-term, and 59.4% of pre-term births were also low birth weight). A deeper analysis of low birth weights and pre-term births is covered in Chapter 4. Each model was adjusted for location, age of the mother at December 31st, 2005 (modelled with a restricted cubic spline with 3 knots), age of the mother at the birth of the child (linear term only), the presence of the outcome in the testing period and testing history or status. Figure 2.30 shows the effect of each of the location effects in both of the models. In general, pregnancies to women from lower socioeconomic and/or more remote areas were more likely to experience either of the adverse outcomes. Figure 2.31 and Figure 2.32 shows the impact of age on the chances of both

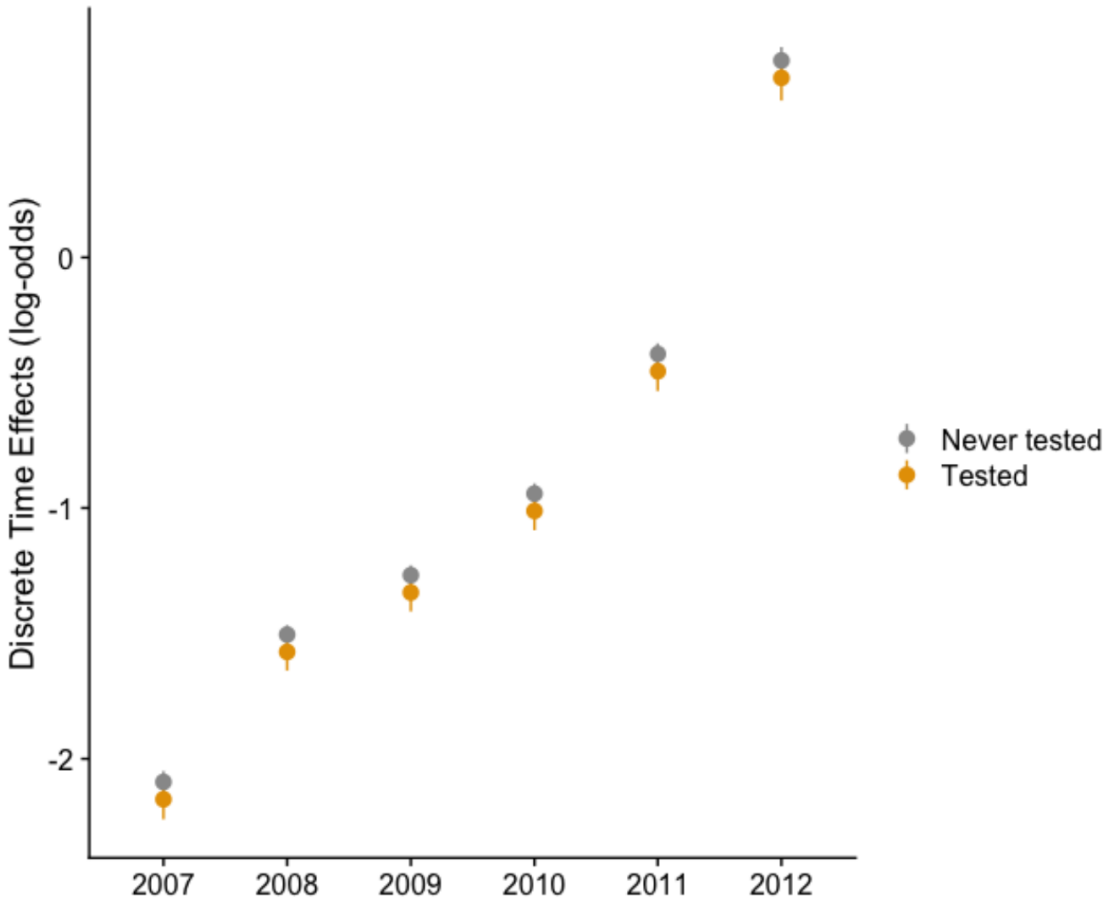

Figure 2.22: **Log-odds of discrete-time effects of exposure to testing on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.**

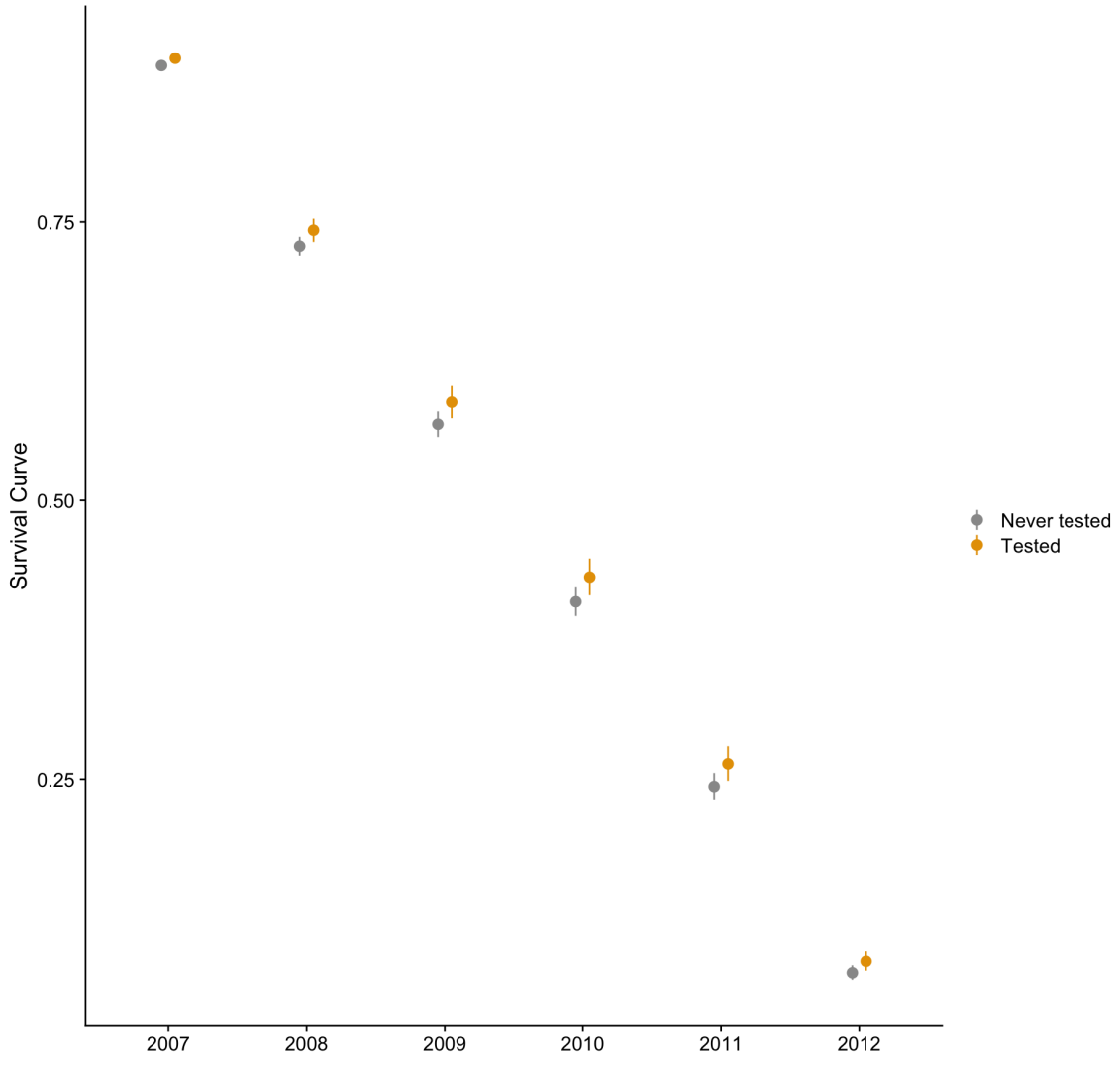

Figure 2.23: **The survival curve of exposure to testing on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.**

Figure 2.24: **Log-odds of discrete-time effects of exposure to testing on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.** This model has an interaction between time and testing status specified.

Figure 2.25: **The survival curve of exposure to testing on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.** This model has an interaction between time and testing status specified.

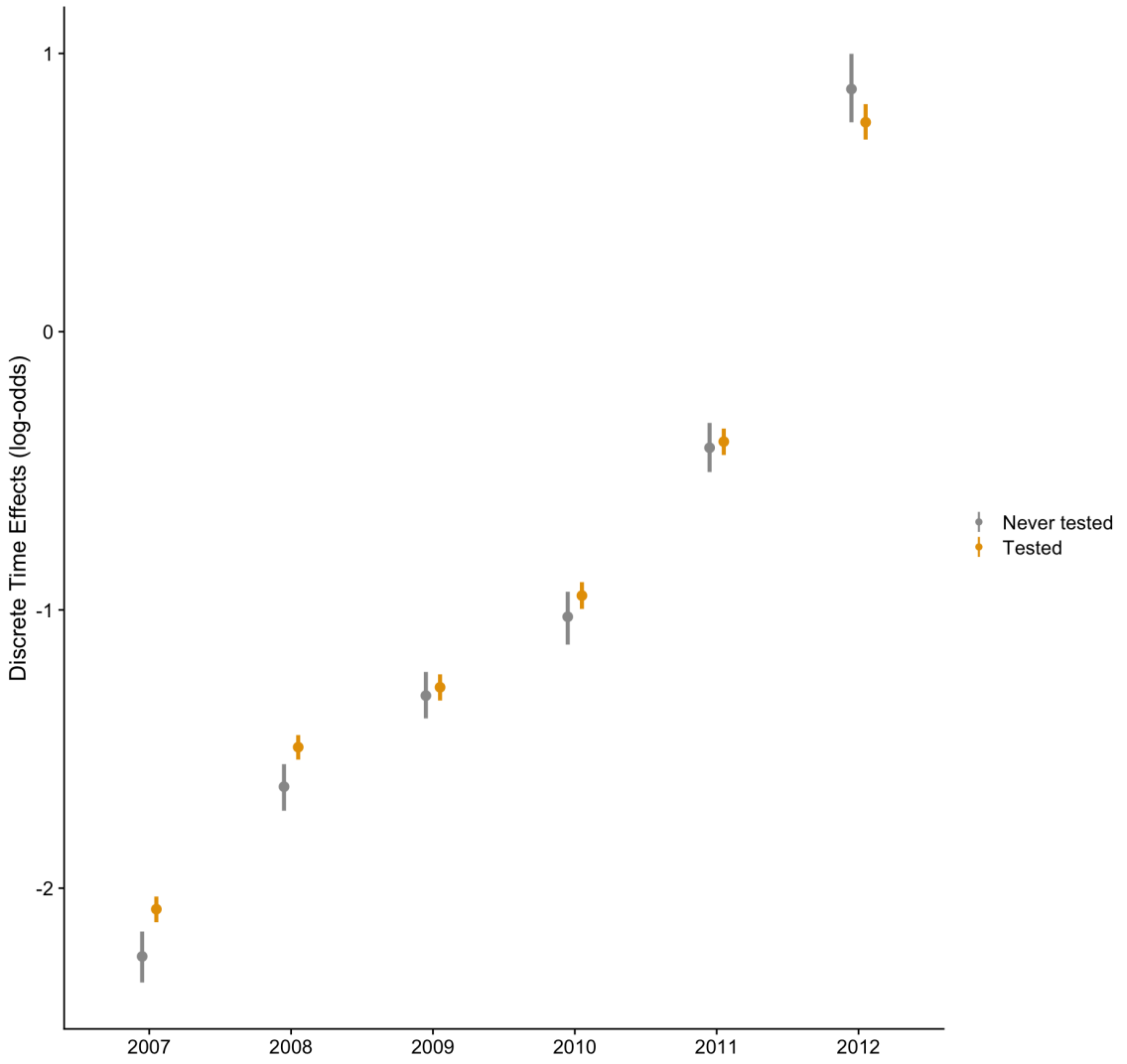

Figure 2.26: **Log-odds of discrete-time effects of exposure to testing on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.** This model has an interaction between time and testing status specified, and the exposure effect is allowed to vary across postcodes.

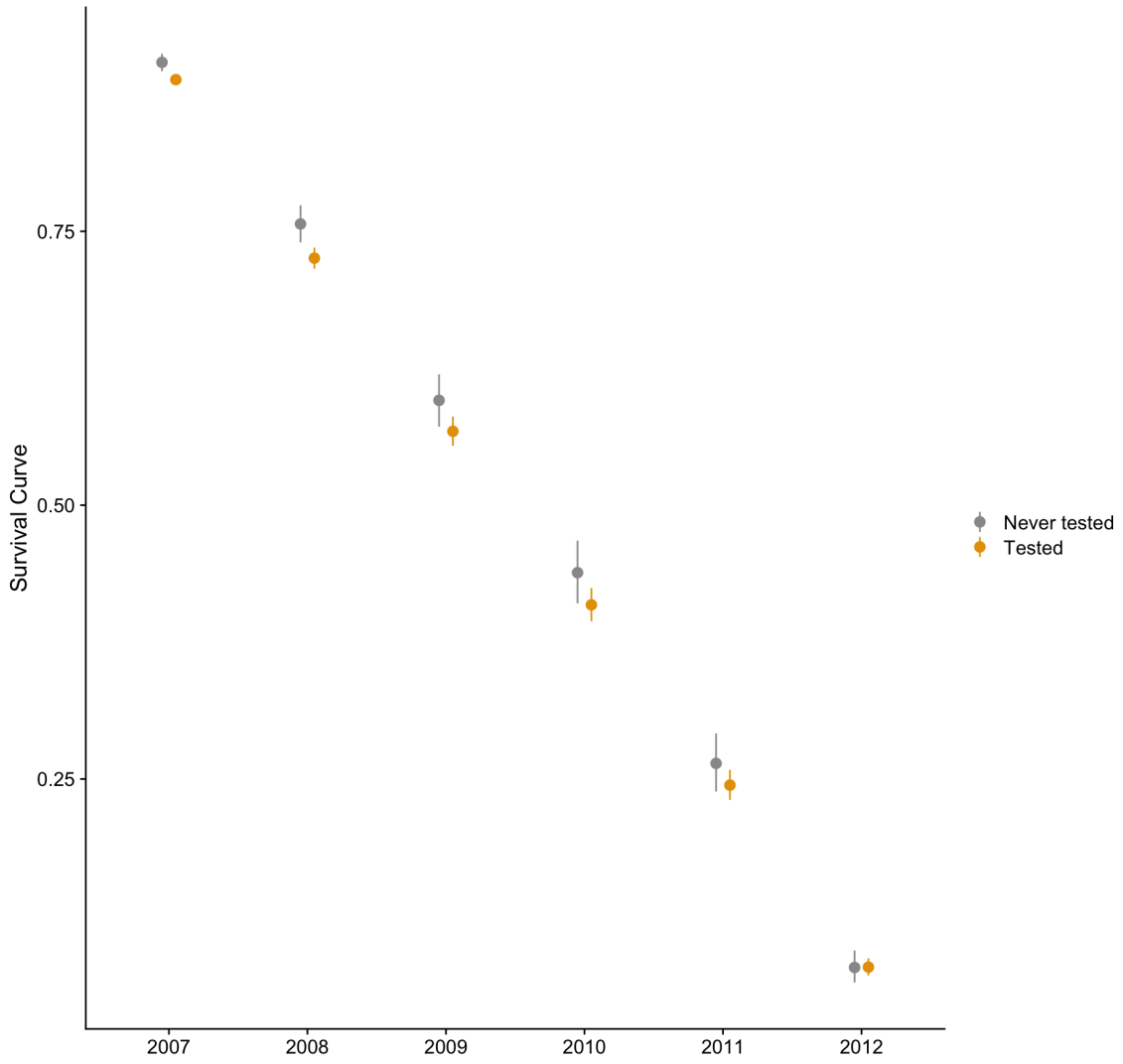

Figure 2.27: **The survival curve of exposure to testing on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.** This model has an interaction between time and testing status specified, and the exposure effect is allowed to vary across postcodes.

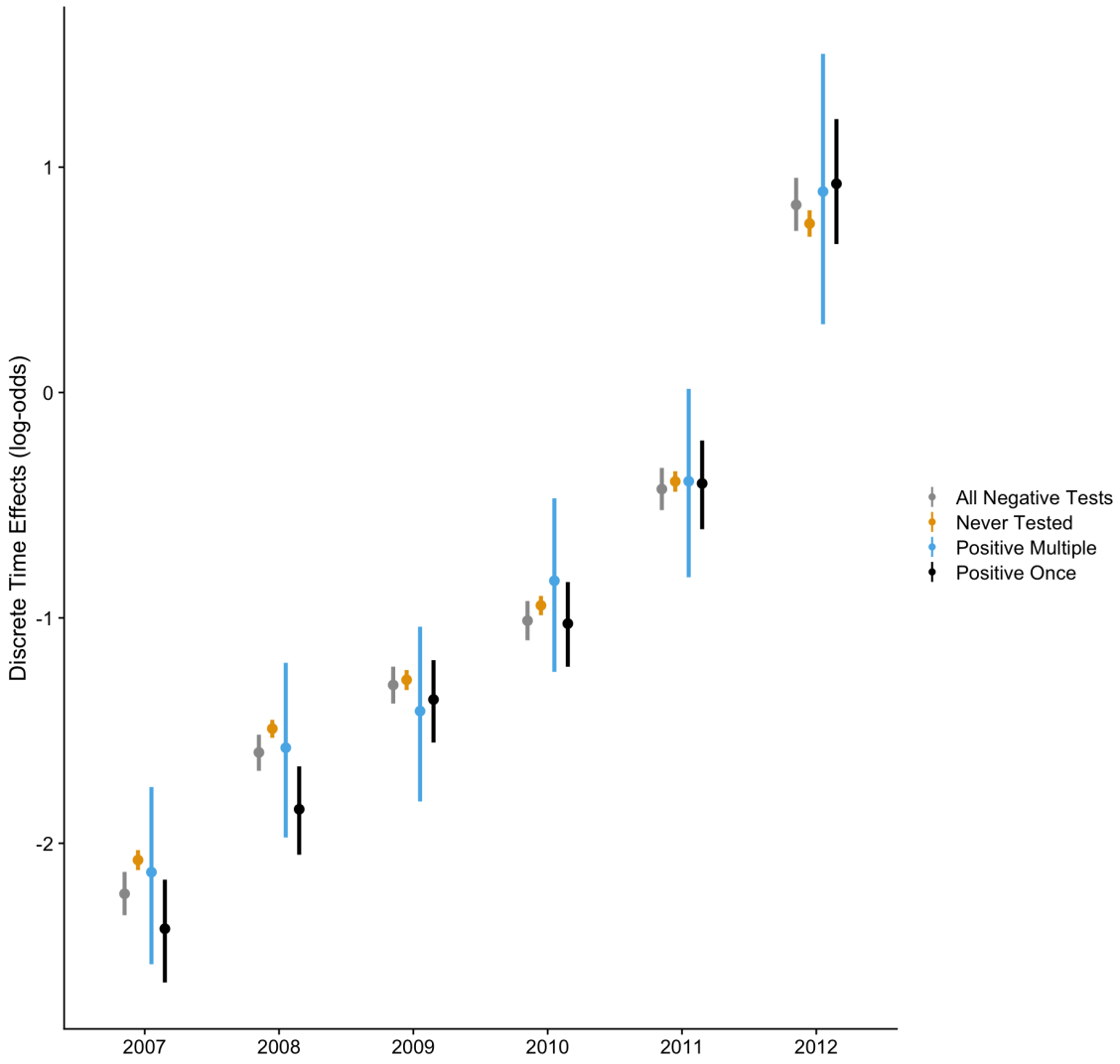

Figure 2.28: **Log-odds of discrete-time effects of exposure to testing, by diagnosis on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.**

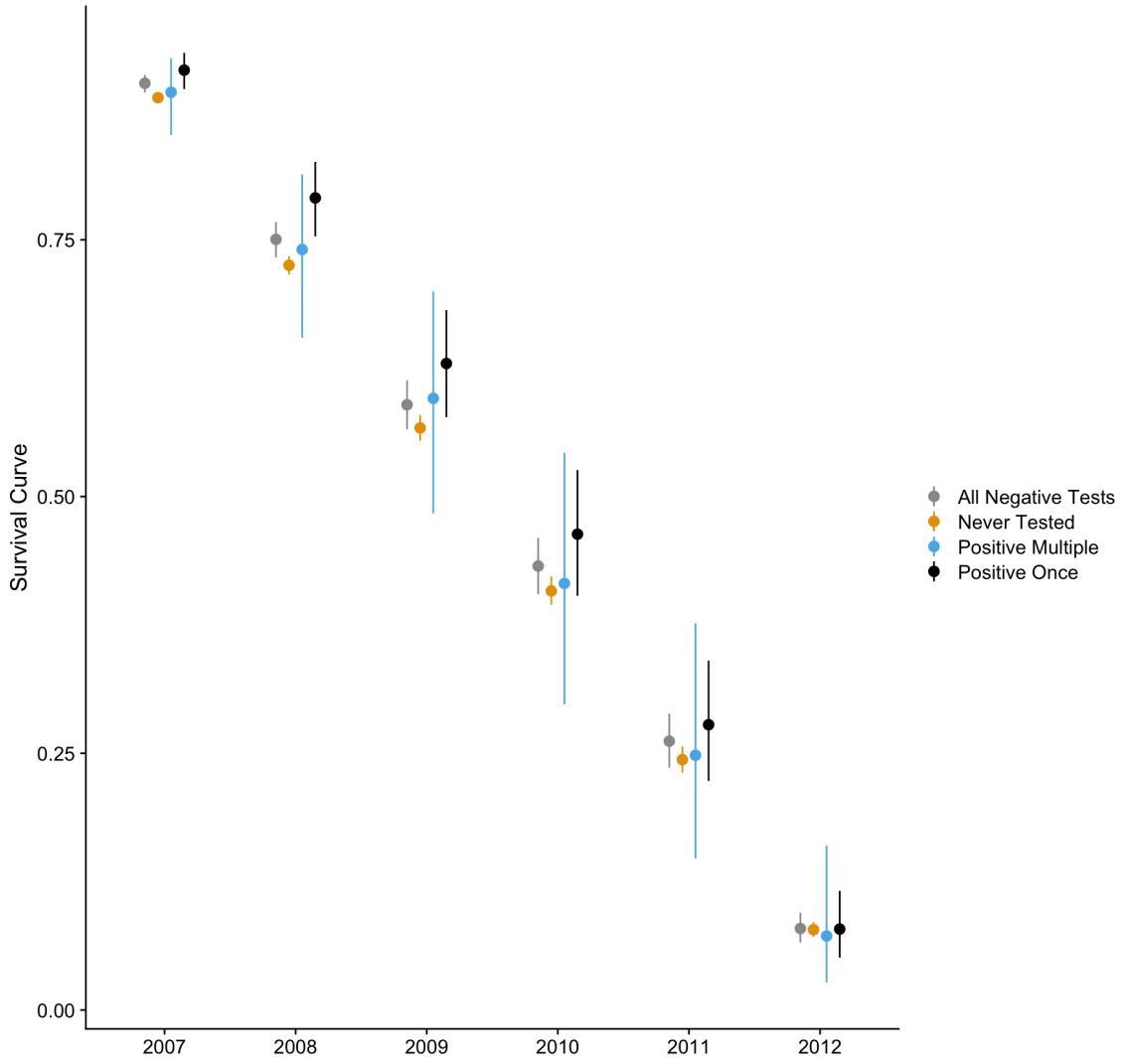

Figure 2.29: **The survival curve of exposure to testing, by diagnosis on pregnancy across years from 2007-2012, for women aged 15-19 at 2005.**

	aOR[a] (95% CI)		aOR (95% CI)
Pre-Term Birth			
		Multiple Positive Tests (n = 290)	1.21 (0.81, 1.81)
Exposed[b] (n = 8,144)	1.28 (1.18, 1.39)	Single Positive Test (n = 1,202)	1.37 (1.14, 1.66)
		All Negative Tests (n = 6,652)	1.27 (1.16, 1.39)
Low Birth Weight			
		Multiple Positive Tests	2.61 (1.85, 3.69)
Exposed	1.35 (1.23, 1.48)	Single Positive Test	1.50 (1.23, 1.84)
		All Negative Tests	1.29 (1.17, 1.43)

Table 2.11: **Perinatal outcomes associated with a history of STI testing and test results.**

[a] Models adjusted for presence of the outcome (pre-term birth or low birth-weight) in the testing period, a cubic spline of age (at 2005) with 3 knots, mother's age at birth and location.

[b] Any history of testing for chlamydia/gonorrhea.

outcomes. Women born later in time (younger at 2005) were less likely to have pre term pregnancies or low birth weight births, and women who were younger when giving birth (independently of the year of their birth) were also less likely to experience adverse perinatal outcomes. Women with a pre term birth in the testing window were much more likely to have another pre term birth in the outcome window (aOR 3.87, 95% CI 3.39 - 4.41), and similarly women with a previous low birth weight baby were much more likely to have another in the outcome window (aOR 4.17, 95% CI 3.60, 4.83). The women who had a history of chlamydia and/or gonorrhea testing were more likely to have adverse perinatal outcomes such as low birth-weight and pre-term birth (Table 2.11; low birth-weight aOR 1.35; 95% CI 1.23–1.48, pre-term birth aOR 1.28; 95% CI 1.18–1.39). Women with multiple positive tests had significantly higher odds of the low birth-weight outcome compared to women with single positive tests and all negative tests (multiple positive tests aOR 2.61; 95% CI 1.85–3.69, single positive test 1.50; 95% CI 1.23–1.84, all negative tests aOR 1.29; 95% CI 1.17– 1.43). The odds of pre-term birth were comparable between all testing groups.

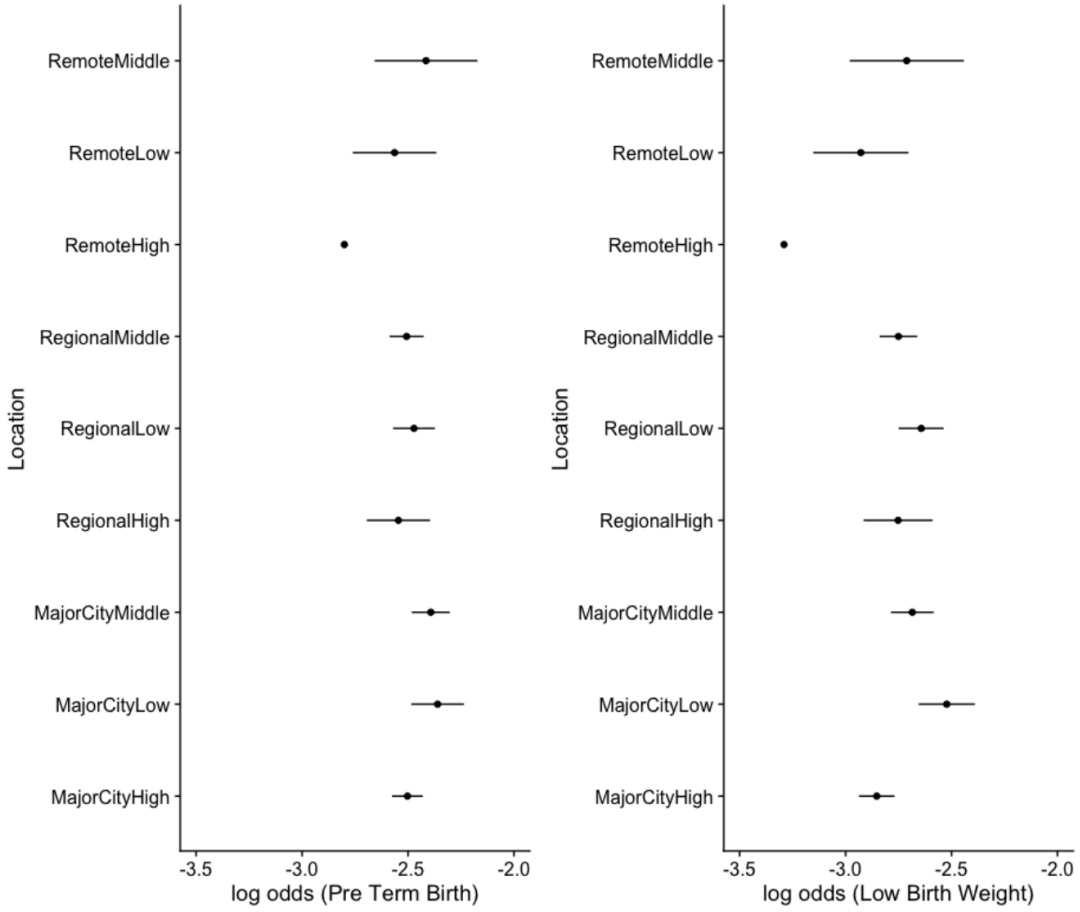

Figure 2.30: **Covariates of the perinatal models and association with each perinatal outcome.** Compares location effects for pre term birth and low birth weight. The effects shown are from the models that use testing history as the key exposure variable.

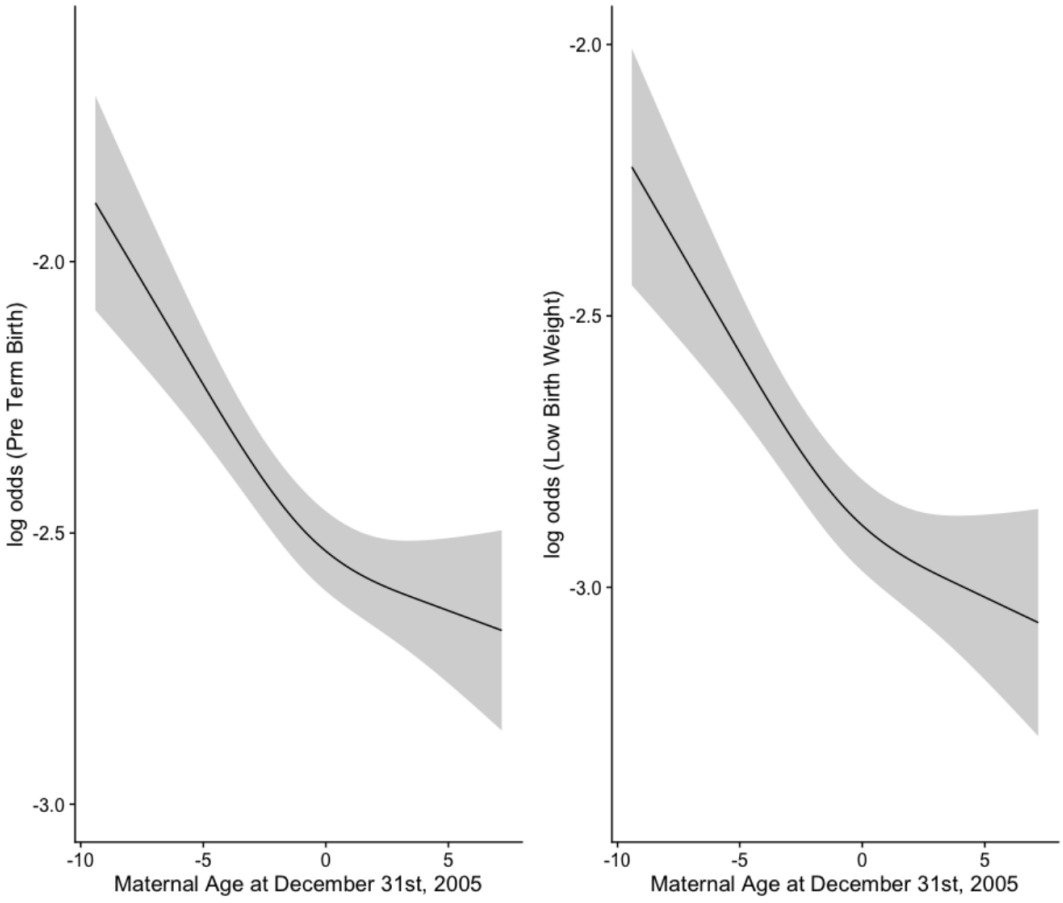

Figure 2.31: **Covariates of the perinatal models and association with each perinatal outcome.** Shows the effect of maternal age (at Dec 31, 2005) on pre term birth and low birth weight. The effects shown are from the models that use testing history as the key exposure variable.

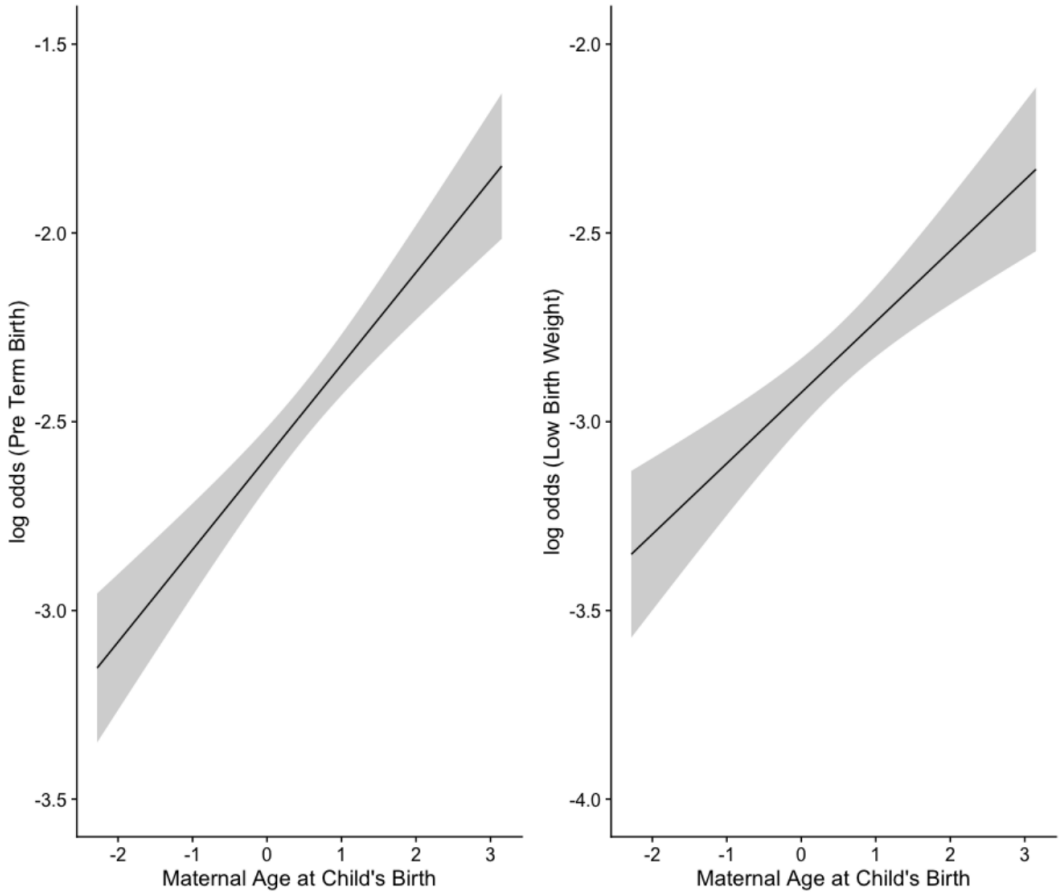

Figure 2.32: **Covariates of the perinatal models and association with each perinatal outcome.** Shows the effect of maternal age (at birth of their child) on pre term birth and low birth weight. The effects shown are from the models that use testing history as the key exposure variable.

2.4 Discussion

In a study of 132 962 women in Queensland, Australia, we have measured the reproductive outcomes for women with a history of testing for chlamydia and gonorrhea. We have investigated how future pregnancies depend on the presence or absence of a testing history, the number of tests, the diagnosis of these tests and the infection. We have also shown that the association between testing history and odds of future pregnancy is dependent on the age at which women are exposed to a test, and if they had any pregnancy during the same time frame as their testing history. This is the first large study using data-linkage to test an association of any recorded subsequent pregnancy with a history of testing for these two STIs.

A strength of our study is the large sample size, an order of magnitude larger than the only other study to look at subsequent pregnancy as an outcome measure [5], and is a similar order of magnitude to studies that have established the association of chlamydial testing history with PID [21, 11, 27, 6] and tubal factor infertility [26]. A further strength of our study is the inclusion of both chlamydia and gonorrhea, and an unexposed group, unlike the only prior study to test for the association with infertility by measuring subsequent pregnancies. Women in the exposed group had reduced odds of pregnancy compared to the unexposed group, when they had no prior pregnancy in the testing period, and they were aged within the middle of the distribution of ages at the end of the testing period. For older and younger ages, and for women with a prior pregnancy in the testing period, there was no difference between exposed and unexposed groups.

There are limitations of our study that should be considered. The time-frame of the cohort did not cover full reproductive years for the women, although it did extend beyond the median age of first birth in QLD in 2013 (28.5) [3] as the cohort has a median of age 31 (IQR 28–34) at study completion. There were 50 185 notifications of chlamydial and gonococcal infections in Queensland from 2000 to 2005 (including all ages and genders) [16], so our data represents $\sim$ 12% of all infections in this time period for this state (numbers for each age group and gender were not available for exact comparison). We were unable to account for people who

had moved out of the study area prior to pregnancies although 2005 census data reported limited migration out of the state so we expect this to have only a small impact [2]. We have adjusted for location and socioeconomic status, however, these factors still may be over-represented in the exposed (tested) group relative to the unexposed group. Relying solely on pathology data and the few associated demographic details means we have no information regarding behavioural and lifestyle predictors of those undergoing STI testing, which risks further confounding of results and difficult interpretation. Further, we were unable to assess the physicians' decision-making process when identifying women for testing, and therefore not able to account for the significance of provider-initiated testing as opposed to patient-initiated testing. It is also possible that women recorded in the perinatal database (which covers both public and private hospitals) may have had chlamydia and gonorrhea tests not recorded in the AUSLAB dataset (government funded testing only). The unexposed group selection was for women who had full blood count test done but no chlamydia and gonorrhea test recorded in AUSLAB. Without knowledge as to why these women were tested, this raises the possibility of further selection bias. The study data did not include dates of tests, or all incidences of negative tests, preventing a temporal analysis that may have provided a richer understanding of the findings.

A previous analysis of population characteristics has established that women who access testing services for chlamydia have lower condom usage and higher numbers of sexual partners, and that these characteristics were more likely to result in chlamydia positivity for women with a testing history compared to a general population [28]. Our findings suggest that STI testing may be indicative of risk behaviours that may predispose individuals to infection and consequently greater adverse outcomes. Care must be taken in interpreting our results to determine the exact reproductive burden of these two STIs, as our definition of exposure implies that women in the exposed group are more sexually active and potentially more likely to have a pregnancy than women in the unexposed group. Women with a pregnancy in the testing period had overall increased chances of having a pregnancy in the outcome window, and this effect was most pronounced for the exposed group. This observation further implies that the women in the exposed group are more

sexually active, and therefore have increased chances of further infections, but also greater chance of further pregnancy. The youngest women in our data had comparable odds of pregnancy between exposed and unexposed groups, however older women in the exposed group had reduced odds. Similarly, younger women with a single positive test had higher chances of pregnancy compared to women with all negative tests. However, there was no distinction between the two for older women. This would suggest that the reproductive burden of chlamydia and gonorrhea increases with age, or with potential further infections that occur over time. Women with multiple positive tests, women with both chlamydia and gonorrhea infections, and women with a testing history that had a pregnancy in the testing window were more likely to have a pregnancy in the outcome window. However, it is important to consider these results in the light of the known differences in health outcomes and STI incidences for Indigenous Australians. The data on Indigenous identity was not available for the unexposed group who did not have a pregnancy, although we can see that with this limited data that there were more Indigenous women in the exposed group. We note that this is likely to be a confounding factor that has particularly contributed to the increased odds of pregnancy results for multiple infections, especially as indigenous women are more likely to test positive to chlamydia and gonorrhea and more likely to have younger pregnancies [14, 4]. The postcode level variation presented from the hierarchical models likely represents shared demographic or environmental factors shared between residents in a similar location, that are not explained by SES effects (as these have been adjusted for already).

Women in the exposed group had increased odds of pre-term births and low birth-weight outcomes. Further, women with a single positive test had odds ratio of both outcomes that were greater than the comparable estimates for women with all negative tests. Although the 95% Confidence Intervals do overlap (Table 2.11), this supports the increased reproductive burden of a positive test diagnosis, in comparison to only negative tests. Location and socioeconomic status was adjusted for in these models, which suggests that the impact of a positive test on perinatal outcomes is due to pathogen and/or host specific factors, rather than being solely caused by demographic factors. The findings support previous findings using outcomes of PID and tubal factor infertility [8, 9, 10, 25, 27].

2.5 Conclusion

We report a relationship between a history of chlamydia or gonorrhea testing history in women with reduced odds of having a subsequent pregnancy for some groups of the population, but not for others, indicating that these two infections could represent a reproductive health burden for some women. It is not clear, from the results presented in this chapter and previous literature measuring burden using PID, whether the association between testing history and reproductive sequelae is attributable to undetected infections that cause tubal damage, or if demographic, behavioural and biological attributes of women with a testing history influences pregnancy outcomes independently of the presence of STIs. Following chapters will utilise further statistical techniques to better elucidate the confounding characteristics of women with a testing history. We will then use a within-host mathematical model of the host-pathogen response to further investigate one process by which STIs could cause reproductive damage, and the potential moderating factors of that process.

2.6 Acknowledgements

We acknowledge the generous collection and provision of data from Pathology Queensland facilitated by the Clinical Information Systems Support Unit, Health Support Queensland and Queensland Health Statistics Unit that were crucial for this study. Deepak Regmi, Database Administrator, Clinical Information Systems and Support Unit, Queensland Health.

References

[1] Standard. 2006. URL: https://www.who.int/childgrowth/standards/technical_report/en/.

[2] *3101.0 - Australian Demographic Statistics, Dec 2005*. Web Page. 2006. URL: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/3101.0Main+Features1Dec%5C%202005.

[3] *3301.0 - Births, Australia, 2013*. Government Document. 2013. URL: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/3301.0Main+Features12013.

[4] *A multivariate approach to the disparity in perinatal outcomes between indigenous and non-indigenous women*. Government Document. 2014. URL: https://www.health.qld.gov.au/__data/assets/pdf_file/0022/370318/indigenous-peridisparity.pdf.

[5] I. J. Bakken et al. "Births and ectopic pregnancies in a large cohort of women tested for Chlamydia trachomatis". In: *Sex Transm Dis* 34.10 (2007), pp. 739–43. ISSN: 0148-5717 (Print) 0148-5717. DOI: 10.1097/01.olq.0000261326.65503.f6.

[6] Inger J Bakken and Sara Ghaderi. "Incidence of pelvic inflammatory disease in a large cohort of women tested for Chlamydia trachomatis: a historical follow-up study". In: *BMC Infectious Diseases* 9.1 (2009), p. 130. ISSN: 1471-2334. DOI: 10.1186/1471-2334-9-130. URL: https://dx.doi.org/10.1186/1471-2334-9-130.

[7] Rachael E. L. Barlow et al. "The prevalence of Chlamydia trachomatis in fresh tissue specimens from patients with ectopic pregnancy or tubal factor infertility as determined by PCR and in-situ hybridisation". In: *Journal of Medical Microbiology* 50.10 (2001), pp. 902–908. ISSN: 0022-2615. DOI: 10.1099/0022-1317-50-10-902.

[8] B. Davies, K. Turner, and H. Ward. "Risk of pelvic inflammatory disease after Chlamydia infection in a prospective cohort of sex workers". In: *Sex Transm Dis* 40.3 (2013), pp. 230–4. ISSN: 0148-5717. DOI: 10.1097/OLQ.0b013e31827b9d75.

[9] B. Davies et al. "Heterogeneity in risk of pelvic inflammatory diseases after chlamydia infection: a population-based study in Manitoba, Canada". In: *J Infect Dis* 210 Suppl 2.Suppl 2 (2014), S549–55. ISSN: 0022-1899 (Print) 0022-1899. DOI: 10.1093/infdis/jiu483.

[10] B. Davies et al. "Pelvic inflammatory disease risk following negative results from chlamydia nucleic acid amplification tests (NAATs) versus non-NAATs

in Denmark: A retrospective cohort". In: *PLoS Med* 15.1 (2018), e1002483. ISSN: 1549-1277 (Print) 1549-1277. DOI: 10.1371/journal.pmed.1002483.

[11] Bethan Davies et al. "Risk of reproductive complications following chlamydia testing: a population-based retrospective cohort study in Denmark". In: *The Lancet Infectious Diseases* 16.9 (2016), pp. 1057–1064. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(16)30092-5. URL: https://doi.org/10.1016/S1473-3099(16)30092-5.

[12] J. L. Goller et al. "Trends in diagnosis of pelvic inflammatory disease in an Australian sexual health clinic, 2002-16: before and after clinical audit feedback". In: *Sex Health* 16.3 (2019), pp. 247–253. ISSN: 1448-5028. DOI: 10.1071/sh18119.

[13] Jane L Goller et al. "Rates of pelvic inflammatory disease and ectopic pregnancy in Australia, 2009–2014: ecological analysis of hospital data". In: *Sexually Transmitted Infections* 94.7 (2018), pp. 534–541. ISSN: 1368-4973. DOI: 10.1136/sextrans-2017-053423.

[14] S. Graham et al. "Epidemiology of chlamydia and gonorrhoea among Indigenous and non-Indigenous Australians, 2000-2009". In: *Med J Aust* 197.11 (2012), pp. 642–6. ISSN: 0025-729x. DOI: 10.5694/mja12.10163.

[15] Frank E. Harrell. *Regression Modeling Strategies*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387952322.

[16] Australian Institute of Health and Welfare. *Rural and remote health*. Report. AIHW, 2019. URL: https://www.aihw.gov.au/reports/rural-remote-australians/rural-remote-health.

[17] Christine L. Heumann et al. "Adverse Birth Outcomes and Maternal Neisseria gonorrhoeae Infection". In: *Sexually Transmitted Diseases* 44.5 (2017), pp. 266–271. ISSN: 0148-5717. DOI: 10.1097/olq.0000000000000592. URL: https://dx.doi.org/10.1097/olq.0000000000000592.

[18] S. D. Hillis et al. "Recurrent chlamydial infections increase the risks of hospitalization for ectopic pregnancy and pelvic inflammatory disease". In: *Am J Obstet Gynecol* 176.1 Pt 1 (1997), pp. 103–7. ISSN: 0002-9378 (Print) 0002-9378. DOI: 10.1016/s0002-9378(97)80020-8.

[19] Jonathan D. Lenz and Joseph P. Dillard. "Pathogenesis of Neisseria gonorrhoeae and the Host Defense in Ascending Infections of Human Fallopian Tube". In: *Frontiers in Immunology* 9 (2018). ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.02710. URL: https://dx.doi.org/10.3389/fimmu.2018.02710.

[20] Bette Liu et al. "Chlamydia and gonorrhoea infections and the risk of adverse obstetric outcomes: a retrospective cohort study". In: *Sexually Transmitted Infections* 89.8 (2013), pp. 672–678. ISSN: 1368-4973. DOI: 10.1136/sextrans-2013-051118. URL: https://dx.doi.org/10.1136/sextrans-2013-051118.

[21] N. Low et al. "Incidence of severe reproductive tract complications associated with diagnosed genital chlamydial infection: the Uppsala Women's Cohort Study". In: *Sex Transm Infect* 82.3 (2006), pp. 212–218.

[22] S. Menon et al. "Human and Pathogen Factors Associated with Chlamydia trachomatis-Related Infertility in Women". In: *Clinical Microbiology Reviews* 28.4 (2015), pp. 969–985. URL: https://cmr.asm.org/content/28/4/969.

[23] Pippa Oakeshott et al. "Randomised controlled trial of screening for Chlamydia trachomatis to prevent pelvic inflammatory disease: the POPI (prevention of pelvic infection) trial". In: *BMJ* 340 (2010). URL: https://www.bmj.com/content/340/bmj.c1642.

[24] Joanne Reekie et al. "Chlamydia trachomatis and the risk of spontaneous preterm birth, babies who are born small for gestational age, and stillbirth: a population-based cohort study". In: *The Lancet Infectious Diseases* 18.4 (2018), pp. 452–460. ISSN: 1473-3099. DOI: 10.1016/s1473-3099(18)30045-8.

[25] Joanne Reekie et al. "Hospitalisations for Pelvic Inflammatory Disease Temporally Related to a Diagnosis of Chlamydia or Gonorrhoea: A Retrospective Cohort Study". In: *PLoS ONE* 9.4 (2014), e94361. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0094361.

[26] Joanne Reekie et al. "Risk of Ectopic Pregnancy and Tubal Infertility Following Gonorrhea and Chlamydia Infections". In: *Clinical Infectious Diseases*

69.9 (2019), pp. 1621–1623. ISSN: 1058-4838. DOI: 10.1093/cid/ciz145. URL: https://dx.doi.org/10.1093/cid/ciz145.

[27] Joanne Reekie et al. "Risk of Pelvic Inflammatory Disease in Relation to Chlamydia and Gonorrhea Testing, Repeat Testing, and Positivity: A Population-Based Cohort Study". In: *Clinical Infectious Diseases* 66.3 (2018), pp. 437–443. ISSN: 1058-4838. DOI: 10.1093/cid/cix769.

[28] J. Riha et al. "Who is being tested by the English National Chlamydia Screening Programme? A comparison with national probability survey data". In: *Sexually Transmitted Infections* 87.4 (2011), pp. 306–311. ISSN: 1368-4973. DOI: 10.1136/sti.2010.047027. URL: https://dx.doi.org/10.1136/sti.2010.047027.

[29] I Simms et al. "Risk factors associated with pelvic inflammatory disease". In: *Sexually Transmitted Infections* 82.6 (2006), pp. 452–457. ISSN: 1368-4973. DOI: 10.1136/sti.2005.019539. URL: https://dx.doi.org/10.1136/sti.2005.019539.

[30] K. Swasdio et al. "The association of Chlamydia trachomatis/gonococcal infection and tubal factor infertility". In: *J Obstet Gynaecol Res* 22.4 (1996), pp. 331–40. ISSN: 1341-8076 (Print) 1341-8076. DOI: 10.1111/j.1447-0756.1996.tb00985.x.

[31] Xavier Troussard et al. "Full blood count normal reference values for adults in France". In: *Journal of Clinical Pathology* 67.4 (2014), pp. 341–344. ISSN: 0021-9746. DOI: 10.1136/jclinpath-2013-201687.

[32] Lars Westrom et al. "Pelvic Inflammatory Disease and Fertility". In: *Sexually Transmitted Diseases* 19.4 (1992), pp. 185–192.

3

Using multiple imputation to correct for binary measurement error in critical confounders, with application to a data linkage study on sexually transmitted infections and pregnancy

3.1 Introduction

Key exposure variable or covariates can often be measured with error in observational data studies. In the case of the gaussian linear model, introducing measurement error into a predictor will attenuate estimates towards the null. When the outcome is binary, and a logistic regression model is used to analyse a particular data set, measurement error may introduce bias into estimates when not appropriately accounted for [23]. In the case of additive error, the estimate of a parameter associated with a mis-measured variable will usually attenuate towards the null. In the case where the mis-measured covariate is discrete and the error is of the misclassification type, the bias may be away from the null if the misclassification is differential (outcome dependent), or if there are confounding covariates [7].

There have been many methods developed to account for misclassification

measurement error in logistic regression models. The misclassification SIMEX handles measurement error by specification of the matrix of true and observed variables, and then simulating data with higher misclassification and extrapolating back to the case of no misclassification [13]. Misclassification measurement error in logistic regression has also been studied in other studies [15, 26], however measurement error corrections are frequently underused in epidemiological and medical studies [12, 3, 19]. This is potentially attributable to a lack of easily available software available to investigators.

In some problems, a gold standard validation dataset exists for a subset of the misclassified predictor. This validation data can be used to construct a relationship between the 'true' values in the validation data and the complete but mis-measured variable of interest [5, 6]. Multiple imputation methods can be used to impute plausible values for the missing data, and the models of interest can be estimated and combined on the imputed data.

3.1.1 Motivating problem

Population level assessments have associated pelvic inflammatory disease (PID) with exposure to Chlamydia trachomatis (chlamydia) and Neisseria gonorrhoeae (gonorrhea), and PID in turn with tubal factor infertility and ectopic pregnancy [24, 14, 9, 18]. Sexually transmitted infections may have obstetric and perinatal complications [17].

Previous studies have evaluated the reproductive burden of the pathogens using subsequent pregnancy as a measure of fertility [2, 4]. A study of a Norwegian population (n = 20762) found no difference in the odds of a pregnancy associated with a positive chlamydial testing result. However, the study provided no comparison to an unexposed group who were not tested for Chlamydia. A later of study of women in QLD, Australia (n = 132962) found significant differences in pregnancy between women who had tested for chlamydia and/or gonorrhea compared to women with no testing history, irrespective of the outcome of the testing history (all negative tests, one positive test, multiple positive tests). Differences in pregnancy

outcomes were found to vary by testing status and age, however no global differences in subsequent pregnancies were found between women with only negative tests, and women with at least one positive test. Women with multiple positive were overall more likely to have a pregnancy, which contradicts the findings around multiple infections and increased risk of PID.

Indigenous people in Australia have higher notification rates for both chlamydia and gonorrhea [11] and have higher rates of hospitalisation for PID [21], although the relative risks of ectopic pregnancy and tubal factor infertility do not necessarily differ [20]. Indigenous women have a higher fertility rate at younger ages than non-Indigenous women [1]. An analysis of subsequent pregnancy given exposure to chlamydial and gonococcal testing including Indigenous status has not previously been possible without accounting for the differences in Indigenous identification between testing and perinatal records. Additionally, in data linkage studies distinct sources are used to collect data for public health and epidemiological analysis, it is likely common to have different priorities and therefore different qualities in the rigour of entry and collection of some variable data. We have used a multiple imputation approach to account for situations where the Indigenous identity of women was collected differently and use this to better evaluate the impact of chlamydial and gonococcal testing on subsequent pregnancy.

3.2 Methods

3.2.1 Data collection and misclassification errors

We use the data derived in Chapter 2 for the analysis in this Chapter, which was obtained from AUSLAB on group of women with exposure to testing for chlamydia or gonorrhea (or both) between 01/01/2000 and 31/12/2005. Testing data was linked to perinatal data from birth entries in QLD between 01/01/2000 and 30/06/2013. Data linkage used deterministic and probabilistic methods, and women where no linkage could be made between testing and perinatal datasets were assumed to have no births during the study time frame.

The data used for this chapter was for women with a history of testing only. The final data had 24,702 observations of women who tested positive or negative for chlamydia and/or gonorrhea in the testing window. Indigenous status was recorded for all women in the testing dataset, and women who gave birth were again recorded for Indigenous status in the perinatal dataset.

3.2.2 Model Building

Models in this chapter are estimated using a maximum likelihood approach for ease of computation; as the structure of the models contain no interaction effects or hierarchical components a Bayesian model would give similar estimates for a longer computation time. Models containing only main effects were used to allow odds ratios between models to be easily compared between imputed and non-imputed data.

The analysis in this chapter was completed using R version 3.6.3. Models and imputation procedures were specified and run using the rms package (v6.0).

3.3 Results

3.3.1 Building the imputation model

	Non-Indigenous (in perinatal data)	Indigenous (in perinatal data)
Non-Indigenous (at testing)	7,806 (72.3%)	1,726 (15.9%)
Indigenous (at testing)	99 (0.1%)	1,170 (10.8%)

Table 3.1: **Comparison of Indigenous status.** Cross-tabulation of Indigenous Identity across testing and perinatal datasets.

Indigenous status data was present for all women in both the testing and perinatal datasets. Our data consisted of women who never identified as Indigenous (7,806 or 72.3% of the cohort), and women who identified at either testing, perinatal or both (Table 3.1). The identification pattern for women was also associated with testing outcomes. Women who never identified as Indigenous were more likely to

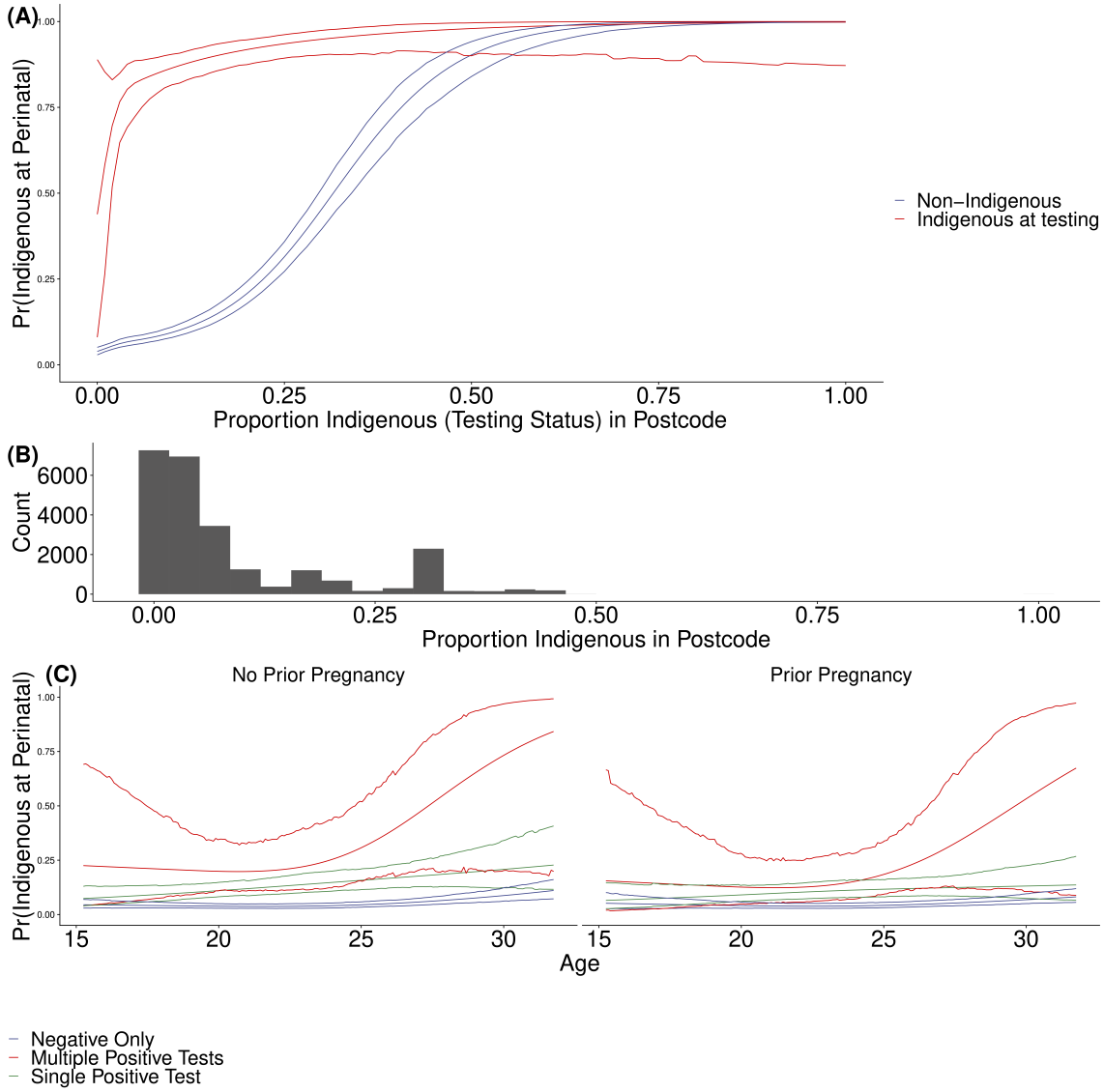

Figure 3.1: **Setting up the imputation model.** (A) Shows the fitted relationship between the proportion of Indigenous women at testing within each postcode, Indigenous status at testing and the prediction of identifying as Indigenous in the perinatal. (B) The histogram of the proportion of Indigenous women in each postcode. (C) The fitted relationship between prior pregnancy, testing status and age, for non-Indigenous women who reside in postcodes with $< 5\%$ of Indigenous women.

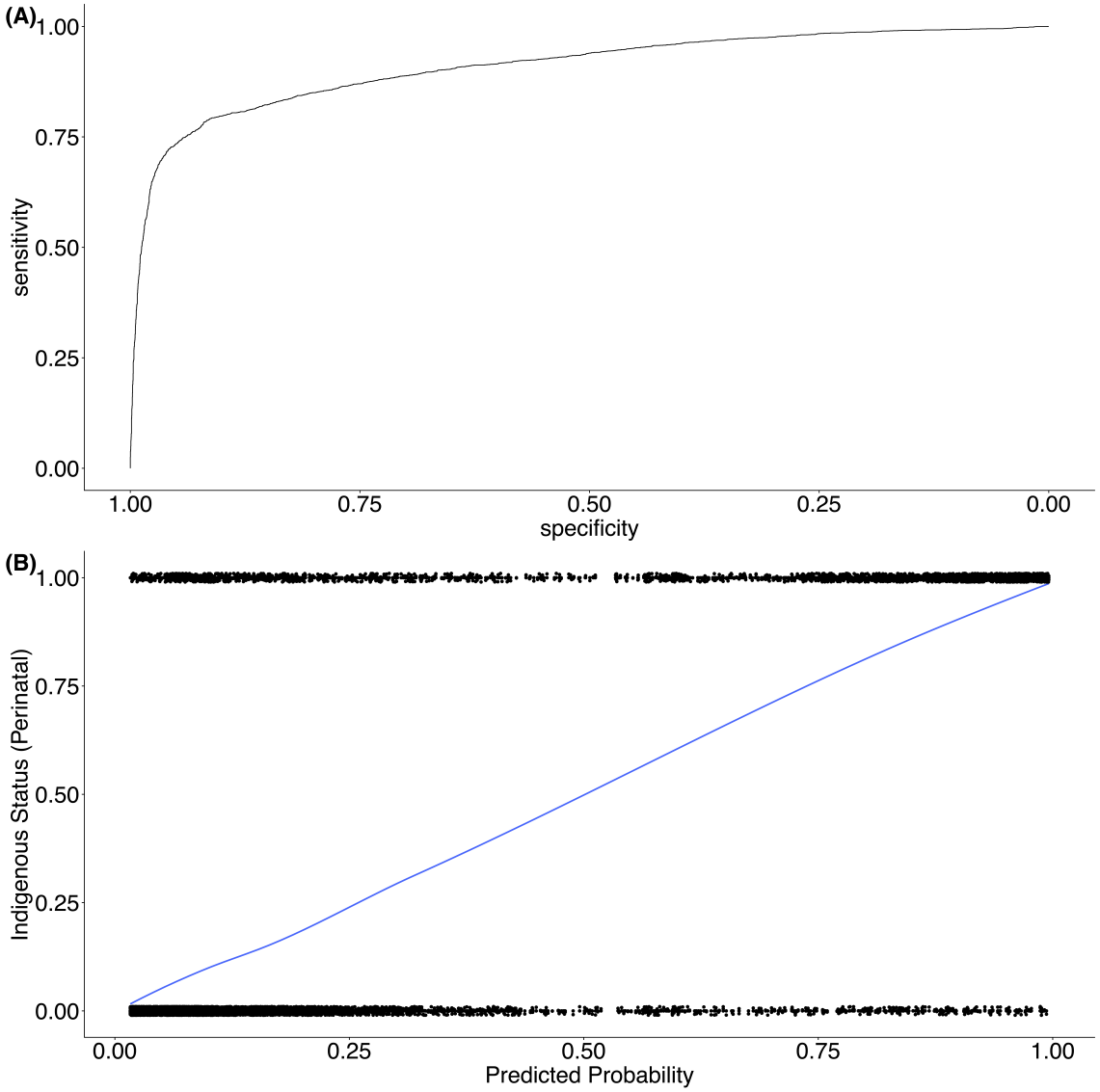

Figure 3.2: **Predictive ability of the Imputation model.** Shows the ROC curve of the trade off between sensitivity and specificity at particular classification thresholds. (B) The calibration between the predicted probability of Indigenous status and the observed outcome of Indigenous status. The blue line is a loess fit of the outcome regressed on the predicted probability.

	All Negative Tests	Single Positive Tests	Multiple Positive Tests
Never Identified as Indigenous	7077	692	37
Indigenous in testing, not in perinatal	72	21	6
Indigenous in perinatal, not in testing	1206	415	105
Always identified as Indigenous	673	309	188

Table 3.2: **Comparison of Indigenous and testing status.** Cross-tabulation of Indigenous Identity across testing and perinatal datasets and testing status for Chalymdia and/or Gonorrhea.

have all negative tests (Table 3.2). Similar distributions of tests were found amongst women who always identified as Indigenous, and those who only identified at perinatal, and not at testing (Figure 3.1). Women who identified only at testing made up a small percentage of the cohort (99 or 0.1%).

Our approach is to treat Indigenous status, recorded in the perinatal data, as a gold-standard measure of a women's preferred identity. This is only recorded for women with a recorded birth in the study time frame, and so it is therefore missing for the remaining dataset. We predict Indigenous at perinatal using Indigenous status at testing, and other key features of the testing data.

The imputation model has a reasonable predictive ability. The AUC of the model is 0.91, implying good discrimination between predicted probabilities and the observed Indigenous outcome, which leads to the trade offs between sensitivity and specificity shown in the ROC curve (Figure **??**(A)). The observed rate of the outcome increases almost linearly with the predicted probability (Figure **??**(B)). The Brier Score of the model was 0.085.

Exploratory analysis indicated that women who did not identify as Indigenous at testing were more likely to identify in the perinatal dataset if they resided in a postcode with a high proportion of Indigenous (recorded at testing) women (Figure 3.1(A), Table 3.3).Women who did not identify as Indigenous at testing, who reside in postcodes with a higher proportion of Indigenous women are more likely to identify as Indigenous as the proportion increases. Women who did iden-

Variable	Coefficient	Standard Error
Intercept	-3.5474	0.4224
Indigenous at testing	3.7807	0.3211
Proportion Indigenous in Postcode - first basis	6.6474	1.5423
Proportion Indigenous in Postcode - second basis	13.9535	6.5996
Multiple positive tests	2.5745	2.0815
Single positive test	1.9514	0.7258
Age (centered) - first basis	0.0782	0.0750
Age (centered) - second basis	-0.1335	0.3990
Age (centered) - third basis	-0.6508	1.7831
Age (centered) - fourth basis	3.7300	2.6702
Prior pregnancy (2005)	-0.9549	0.7633
Major City and Low SES	1.0893	0.2128
Major City and Middle SES	0.4772	0.1997
Regional and High SES	1.1272	0.1883
Regional and Low SES	1.2969	0.1663
Regional and Middle SES	0.9777	0.1587
Remote and Low SES	2.1786	0.1995
Remote and Middle/High SES	1.8015	0.2162
Indigenous at testing and Proportion Indigenous in Postcode - first basis	5.9697	5.1995
Indigenous at testing and Proportion Indigenous in Postcode - second basis	-52.1019	20.4222
Multiple positive tests and Age (centered) - first basis	0.1383	0.3968
Single positive test and Age (centered) - first basis	0.1809	0.1345
Multiple positive tests and Age (centered) - first basis	-1.3064	2.0651
Single positive test and Age (centered) - second basis	-0.6166	0.7310
Multiple positive tests and Age (centered) - third basis	7.4414	9.5108
Single positive test and Age (centered) - third basis	2.0884	3.2997
Multiple positive tests and Age (centered) - fourth basis	-11.9431	15.6664
Single positive test and Age (centered) - fourth basis	-2.3399	5.0045
Prior pregnancy and Age (centered) - first basis	-0.1726	0.1530
Prior pregnancy and Age (centered) - second basis	0.6626	0.7014
Prior pregnancy and Age (centered) - third basis	-1.5768	2.8874
Prior pregnancy and Age (centered) - fourth basis	-0.2799	3.9900

Table 3.3: **Estimates of identifying as Indigenous at perinatal in the outcome window.** Model estimates are maximum likelihood estimates.

tify as Indigenous at testing are equally as likely to identify at perinatal given their postcode, unless that postcode has almost no Indigenous women. However, almost all women resided in postcodes with less than 50% Indigenous, and a majority of women were in postcodes of less than 5% Indigenous women.

Women were more likely to identify in the perinatal data if they had multiple positive tests, and generally in the older age ranges observed in the cohort (Figure 3.1(C), Table 3.3). There were no observable differences between women with and without a pregnancy in the testing period and identifying as Indigenous.

3.3.2 Main effects model to determine the impact of imputation

Variable	No Indigenous Model **Estimate** (**Std. Error**)	Indigenous testing Model **Estimate** (**Std. Error**)	Imputation Model **Estimate** (**Std. Error**)
Intercept	-1.701 (0.0454)	-1.1573 (0.0455)	-1.1555 (0.0456)
Multiple positive tests	0.4279 (0.0865)	0.2751 (0.0887)	0.2373 (0.0890)
Single positive test	0.1349 (0.0404)	0.0990 (0.0408)	0.0299 (0.0419)
Indigenous (none, testing, imputed)	NA	0.4685 (0.0505)	0.5884 (0.0502)
Age (Centered) - First Basis	-0.0309 (0.0095)	-0.0221 (0.009)	-0.0249 (0.0096)
Age (Centered) - Second Basis	-0.0146 (0.0117)	-0.0173 (0.0117)	-0.0235 (0.01178)
Pregnant in the testing window	0.3939 (0.0356)	0.3752 (0.0357)	0.3709 (0.0359)
Major City and Low SES	0.4115 (0.0743)	0.3980 (0.0744)	0.3704 (0.0749)
Major City and Middle SES	0.0942 (0.0542)	0.0918 (0.0542)	0.0846 (0.0543)
Regional and High SES	0.0622 (0.0598)	0.0549 (0.0598)	0.0330 (0.0601)
Regional and Low SES	0.7000 (0.0466)	0.6410 (0.0471)	0.5170 (0.0493)
Regional and Middle SES	0.5163 (0.0397)	0.4926 (0.0398)	0.4538 (0.0402)
Remote and Low SES	0.6874 (0.0550)	0.5726 (0.0556)	0.2690 (0.0654)
Remote and Middle/High SES	0.5472 (0.0798)	0.4739 (0.0805)	0.2774 (0.0839)

Table 3.4: **Estimates of pregnancy in the outcome window, using imputed Indigenous status as a predictor.** Model estimates are maximum likelihood estimates.

	No Indigenous Predictor - Odds Ratio (95% CI)	Indigenous at Testing Predictor - Odds Ratio (95% CI)	Imputed Indigenous Predictor - Odds Ratio (95% CI)
Multiple Positive Tests	1.53 (1.29 - 1.82)	1.32 (1.11 - 1.57)	1.27 (1.06 - 1.51)
Single Positive Tests	1.14 (1.06 - 1.24)	1.10 (1.02 - 1.20)	1.03 (0.95 - 1.12)

Table 3.5: **Adjusted odds ratios of pregnancy in the outcome window.** Compares models with no Indigenous predictor, Indigenous at testing and imputed Indigenous status. Model estimates are maximum likelihood estimates.

The cohort consisted of 24,702 women who had at least one test for Chlamydia and/or Gonorrhea in the time frame of the study (2000-2005). Without adjusting

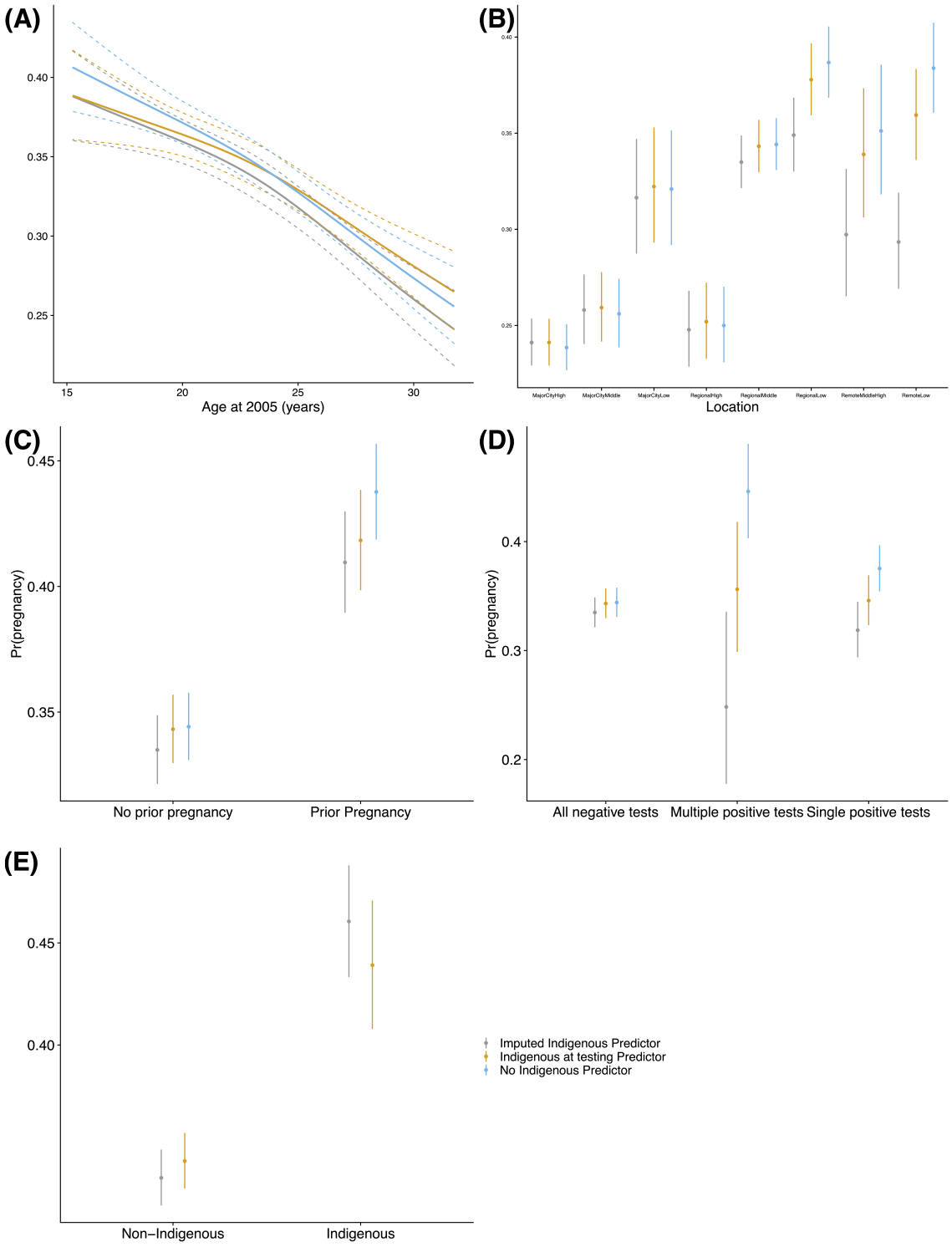

Figure 3.3: **The predicted probability of pregnancy in the outcome window by predictors of the models.** (A) Compares the age relationship between models with no Indigenous predictor, Indigenous at testing and imputed Indigenous status. (B) Compares location effects. (C) Compares prior pregnancy effects. (D) Compares testing effects. (E) Compares Indigenous status. Predicted probabilities are adjusted (where applicable) to non-Indigenous status, major city and high SES locations, average age, negative tests and no prior pregnancy.

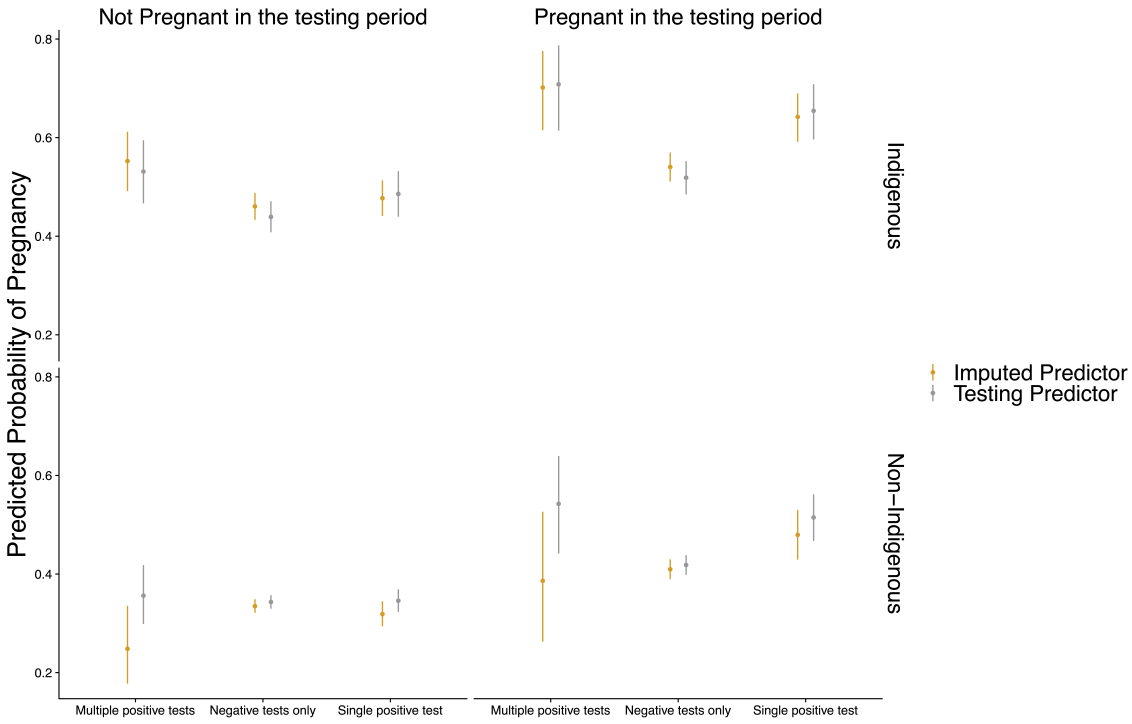

Figure 3.4: **The predicted probability of pregnancy in the outcome window by testing status, prior pregnancy and Indigenous status.** Compares estimates between models with predictors of Indigenous at testing and imputed Indigenous status. Predicted probabilities are adjusted average age and Regional and Middle SES locations.

for Indigenous status, older women were overall less likely to have a pregnancy in the outcome window (Figure 3.3(A), Table 3.4). Women with a prior pregnancy in the testing window had an adjusted odds ratio of 1.48 (95% CI 1.38 - 1.59) of pregnancy in the outcome window (Figure 3.3(C)). Women in more remote and lower SES postcodes had increased odds of pregnancy (Table 3.4).

Women with positive tests had increased odds of pregnancy compared to women with only negative tests. The increase in the odds was higher for women with multiple positive tests (aOR 1.53; 95% CI 1.29 - 1.82) than for women with a single positive test (aOR 1.14; 95% CI 1.06 - 1.24). This effect was observed in the previous chapter for all age ranges.

Once Indigenous status was included as a predictor, using Indigenous status at testing, we found that Indigenous women had increased odds of pregnancy in the outcome window (aOR 1.60; 95% CI 1.45 - 1.76). When using the imputed Indigenous status, the chances of Indigenous women having a pregnancy in the outcome window were estimated to have increased (Figure 3.3(E), Table 3.4).

After accounting for Indigenous status, the magnitude of the positive testing effects were reduced (Figure 3.3(D), Table 3.4, Table 3.5). In particular, the confidence interval for the adjusted odds ratio for women with single positive tests covered 0 after adjusting for imputed Indigenous status (Table 3.5), suggesting that there was no statistical difference between women with single positive tests and women with all negative tests. However, the odds for women with multiple positive tests remained larger than the odds for women with all negative tests.

3.3.3 Using the imputation process to further understand the impact of testing status

In the previous chapter we described the process of deduction and utilising prior inference that we used to derive the models that we described in the results section discussed in Chapter 2. Specifically, the presence of a prior pregnancy in the testing period was interacted with testing status, as we believed a priori that women may have been tested as part of routine pregnancy checks. The presence of a

Variable	Indigenous testing Model **Estimate** (**Std. Error**)	Imputation Model **Estimate** (**Std. Error**)
Intercept	-1.1464 (0.0456)	–1.1451 (0.0458)
Multiple positive tests	0.0572 (0.1320)	-0.4161 (0.2113)
Single positive test	0.0118 (0.0482)	-0.0695 (0.0578)
Indigenous (testing, imputed)	0.4043 (0.0612)	0.5385 (0.0641)
Age (Centered) - First Basis	-0.0220 (0.0096)	-0.0254 (0.0097)
Age (Centered) - Second Basis	-0.0166 (0.0118)	-0.0218 (0.0119)
Pregnant in the testing window	0.3195(0.0387)	0.3184 (0.0390)
Major City and Low SES	0.4032 (0.0744)	0.3785 (0.0747)
Major City and Middle SES	0.0968 (0.0542)	0.0902 (0.0544)
Regional and High SES	0.0582 (0.0598)	0.0377 (0.0600)
Regional and Low SES	0.6479 (0.0472)	0.5250 (0.0500)
Regional and Middle SES	0.4975 (0.0398)	0.4614 (0.0402)
Remote and Low SES	0.5687 (0.0568)	0.2653 (0.0841)
Remote and Middle/High SES	0.4787 (0.0805)	0.02653 (0.0687)
Multiple positive tests and pregnant in the testing window	0.4422 (0.2155)	0.3288 (0.2179)
Single positive test and pregnant in the testing window	0.3768 (0.1082)	0.3594 (0.1090)
Multiple positive tests and Indigenous	0.3124 (0.1809)	0.7734 (0.2420)
Multiple positive tests and Indigenous	0.1757 (0.1127)	0.1211 (0.0974)

Table 3.6: **Estimates of pregnancy in the outcome window, using imputed Indigenous status as a predictor.** Model estimates are maximum likelihood estimates.

	Indigenous at Testing Predictor - Odds Ratio (95% CI)	Imputed Indigenous Predictor - Odds Ratio (95% CI)
Multiple Positive Tests - Non-Indigenous and no prior pregnancy	1.06 (0.82 - 0.37)	0.66 (0.44 - 0.99)
Single Positive Tests - Non-Indigenous and no prior pregnancy	1.011 (0.92 - 1.11)	0.93 (0.83 - 1.04)
Multiple Positive Tests - Non-Indigenous and prior pregnancy	1.65 (1.10 - 2.48)	0.92 (0.53 - 1.60)
Single Positive Tests - Non-Indigenous and prior pregnancy	1.47 (1.21 - 1.80)	1.34 (1.08 - 1.65)
Multiple Positive Tests - Indigenous and no prior pregnancy	1.45 (1.10 - 1.90)	1.43 (1.13 - 1.81)
Single Positive Tests - Indigenous and no prior pregnancy	1.21 (0.98 - 1.48)	1.05 (0.95 - 1.12)
Multiple Positive Tests - Indigenous and prior pregnancy	2.25 (1.46 - 3.48)	1.27 (1.13 - 1.81)
Single Positive Tests - Indigenous and prior pregnancy	1.78 (1.34 - 2.31)	1.05 (0.90 - 1.23)

Table 3.7: **Adjusted odds ratios of pregnancy in the outcome window.** Compares models with Indigenous at testing and imputed Indigenous status. Model estimates are maximum likelihood estimates.

pregnancy prior to the onset of the outcome window also indicates a lower likelihood of experiencing pathology.

In line with that argument, we also believe that the effects of testing status (negative only, positive once, positive multiple times) should be different for Indigenous and non-Indigenous women. This is supported by the evidence in the previous chapter that described how the effects of testing varies across postcode, indicating that demographic factors indicates the presence of differing host and pathogen factors that may influence the progression of an infection to pathology.

The previous subsection used an under-specified main effects model to investigate the impact of the imputation process. The imputation process demonstrates a clear distinction in the estimates of the key odds ratios of interest, so we will use this process to examine the odds ratios when cohorted by key variables of prior pregnancy and Indigenous status.

The imputation process, as in the previous set of results, did not significantly alter the estimates associated with age, the main effect of prior pregnancy and location effects (Table 3.6).

The estimates and adjusted odds ratios for women with single and multiple positive tests were significantly different across the groups of prior pregnancy and Indigenous status, and between models using different Indigenous predictors (Table 3.6, Table 3.7, Figure 3.4).

The odds ratios and associated 95% confidence intervals were above unity for women with single positive tests and with a prior pregnancy and/or who identified as Indigenous at testing. The equivalent odds ratios attenuated towards unity when the imputed Indigenous status was incorporated. The confidence intervals for the single positive test effect, when using the imputed Indigenous status variable, covered unity except for non-Indigenous women with prior pregnancies, which corresponds to our prior belief about the confounding of testing and pregnancy. Interestingly, this is not the case for Indigenous women.

There was a clear distinction between the effects of multiple positive tests

between Indigenous and non-Indigenous women, once the imputed Indigenous status was incorporated. The adjusted odds ratio for non-Indigenous women with no prior pregnancies was 0.66 (95% CI 0.44 - 0.99). For Indigenous women, having multiple positive tests indicated an increase in the chances of pregnancy in the outcome window, independently of any pregnancies in the testing window.

3.4 Discussion

Data linkage studies are increasingly important epidemiological and public health tool with the increase usage of medical systems and the need to use large datasets to get better knowledge of population level effects. Largely health management systems are not constructed for the purpose of research analysis. Additionally, data collection, need for data entry into the systems (in this case pathology requests are not directly linked to the clinical record but a separate request and data system) and in some cases data disclosure by individuals all impact on quality of source data.

We encountered these issues in our recent study to attempt to evaluate pregnancies, as indicated by a perinatal birth record, against data indicating exposure and testing history for chlamydia and gonorrohea. In particular, using pathology records, we found that pathology test requests did not require Indigenous status disclosure for FBC, but it was reported (albeit inconsistently) for STI testing requests.

Whilst we are not aware of the specific causes for inconsistency between Indigenous status between records, we found that women who identified as Indigenous in the perinatal records almost always identified (1,170 out of 1,269 identified in this way), whereas women who were identified as non-Indigenous in the testing records were often identified as Indigenous in the perinatal records (1,726 women identified in the perinatal records out of 9,532 non-Indigenous in the testing records). We can speculate a combination of factors. Perinatal disclosure is crucial for the newborn to have a public health record of Indigenous identity, so more likely to be reported and accurately recorded by clinicians. STI clinics know the status of their clients and do not always see the need to report on the pathology request. Individuals at

STI testing status may also not want to disclose Indigenous status to a clinician.

This motivated a need to apply a misclassification correction to our data, to allow adjusting for Indigenous status within models, as the measurement error literature indicates that not applying a correction method introduces bias into the estimates of a logistic regression model [23, 5], and that the bias is not necessarily attenuated towards the null [7]. The perinatal record of Indigenous status acts as a validation record for Indigenous status as recorded in the testing records, which allows for the misclassification correction to be cast as a missing data problem, and for a multiple imputation method to be applied. This approach has previously been demonstrated to have similar performance (in terms of bias, precision and convergence) compared to misclassified and regression calibration analyses [6], but also has ease of use as modern software packages for multiple imputation can be easily applied.

After adjusting for imputed Indigenous status, we found no difference between women with all negative tests and single positive tests. The movement of the odds ratios after imputation suggests that inclusions of further key confounders of pregnancy and testing status may further influence the difference between negative and positive testing groups. Key confounders may include contraceptive use and behavioural factors such as those described in survey data [22]. The progression to pathology may also be influenced by key host and pathogen factors [16, 25], which are investigated in later chapters. However, there may be little risk of reduced pregnancy attributable to a single positive test, as previously noted in the literature [2].

We also found that the difference between women with all negative tests and multiple positive tests was attenuated once the imputed main effects model was used to estimate the key odds ratios. Once key interaction effects were include in the models, the impact of multiple positive tests were found to differentiate strongly between non-Indigenous and Indigenous women. The results for non-Indigenous women indicated a moderately strong association between multiple infections and adverse reproductive outcomes. The relationship between multiple positive tests and tubal factor infertility have been previously been investigated, and there was no significant difference found between women with multiple and single positive

tests for the TFI outcome [9], although multiple positive tests clearly increases the risk of PID [9, 8]. As far as we are aware, this is therefore the first time that the potential risk of multiple positive tests and reduced pregnancy has been identified. This supports the usage of pregnancy as an outcome measure, as the increase in sample size provides the ability to distinguish between multiple competing factors in our analysis. It is not clear why the effect of multiple positive tests may differ for Indigenous women. Indigenous women have higher notification rates for both chlamydia and gonorrhea [11]. The impact of multiple infections may therefore be different if different pathogens are present each time, rather than re-infection with Chlamydia only.

The imputation process described in this chapter could be extended to the hierarchical models discussed in the previous chapter to fully account for the interaction between testing effects and age, and the variation of these effects across postcodes once Indigenous status has been accounted for. Bayesian methods would be needed to appropriately estimate the effects across different postcodes when multiply imputed datasets are used. As indicated in a simulation study, the combining of effects using Rubin's rules across datasets may not lead to appropriate uncertainty intervals when the posterior distribution (or the sampling distribution of the complete-data estimator) is not approximately Gaussian [27]. A more reliable approach would be to stack the posterior distributions, with a large number of imputed datasets, or to have measurement error incorporated as part of the model [10]. We would note that the results in this chapter should be interpreted with this point in mind, although the models used in this chapter are simple in structure. This concern may be more relevant for a more richly specified model structure.

3.5 Conclusion

Overall, we validate that imputation can be used to account for source data inconsistencies in critical variables. We found that when we look solely at exposed groups and adjusted for imputed indigenous status that the adjusted odds ratio for non-Indigenous women with a single positive tests better reflects the prior work on testing status and PID, that women with a negative test should not be more likely

to experience pathology. Our results here support the application of imputation to future data linkage studies as a valid approach to overcome limitations and source data.

References

[1] *3101.0 - Australian Demographic Statistics, Dec 2005*. Web Page. 2006. URL: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/3101.0Main+Features1Dec%5C%202005.

[2] I. J. Bakken et al. "Births and ectopic pregnancies in a large cohort of women tested for Chlamydia trachomatis". In: *Sex Transm Dis* 34.10 (2007), pp. 739–43. ISSN: 0148-5717 (Print) 0148-5717. DOI: 10.1097/01.olq.0000261326.65503.f6.

[3] T. B. Brakenhoff et al. "Measurement error is often neglected in medical literature: a systematic review". In: *J Clin Epidemiol* 98 (2018), pp. 89–97. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2018.02.023.

[4] Torrington Callan et al. "A retrospective cohort study examining STI testing and perinatal records demonstrates reproductive health burden of chlamydia and gonorrhea". In: *Pathogens and Disease* 78.6 (2020). ISSN: 2049-632X. DOI: 10.1093/femspd/ftaa052. URL: https://doi.org/10.1093/femspd/ftaa052.

[5] Raymond J Carroll et al. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006. ISBN: 1420010131.

[6] Stephen R Cole, Haitao Chu, and Sander Greenland. "Multiple-imputation for measurement-error correction". In: *International Journal of Epidemiology* 35.4 (2006), pp. 1074–1081. ISSN: 0300-5771. DOI: 10.1093/ije/dyl097. URL: https://doi.org/10.1093/ije/dyl097.

[7] Ori Davidov, David Faraggi, and Benjamin Reiser. "Misclassification in Logistic Regression with Discrete Covariates". In: *Biometrical Journal* 45.5 (2003), pp. 541–553. ISSN: 0323-3847. DOI: https://doi.org/10.1002/bimj.

200390031. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200390031.

[8] B. Davies et al. "Heterogeneity in risk of pelvic inflammatory diseases after chlamydia infection: a population-based study in Manitoba, Canada". In: *J Infect Dis* 210 Suppl 2.Suppl 2 (2014), S549–55. ISSN: 0022-1899 (Print) 0022-1899. DOI: 10.1093/infdis/jiu483.

[9] Bethan Davies et al. "Risk of reproductive complications following chlamydia testing: a population-based retrospective cohort study in Denmark". In: *The Lancet Infectious Diseases* 16.9 (2016), pp. 1057–1064. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(16)30092-5. URL: https://doi.org/10.1016/S1473-3099(16)30092-5.

[10] Andrew Gelman et al. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[11] S. Graham et al. "Epidemiology of chlamydia and gonorrhoea among Indigenous and non-Indigenous Australians, 2000-2009". In: *Med J Aust* 197.11 (2012), pp. 642–6. ISSN: 0025-729x. DOI: 10.5694/mja12.10163.

[12] Anne M. Jurek et al. "Exposure-measurement error is frequently ignored when interpreting epidemiologic study results". In: *European Journal of Epidemiology* 21.12 (2007), pp. 871–876. ISSN: 0393-2990. DOI: 10.1007/s10654-006-9083-0.

[13] Helmut Küchenhoff, Samuel M. Mwalili, and Emmanuel Lesaffre. "A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX". In: *Biometrics* 62.1 (2006), pp. 85–96. ISSN: 0006-341X. DOI: 10.1111/j.1541-0420.2005.00396.x.

[14] N. Low et al. "Incidence of severe reproductive tract complications associated with diagnosed genital chlamydial infection: the Uppsala Women's Cohort Study". In: *Sex Transm Infect* 82.3 (2006), pp. 212–218.

[15] Robert H. Lyles and Ji Lin. "Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting". In: *Statistics in Medicine* 29.22 (2010), pp. 2297–2309. ISSN: 0277-6715. DOI: 10.1002/sim.3971. URL: https://dx.doi.org/10.1002/sim.3971.

[16] Dann G. Mallet et al. “A Mathematical Model of Chlamydial Infection Incorporating Movement of Chlamydial Particles”. In: *Bulletin of Mathematical Biology* 75.11 (2013), pp. 2257–2270. ISSN: 1522-9602. DOI: 10.1007/s11538-013-9891-9. URL: https://doi.org/10.1007/s11538-013-9891-9.

[17] S. Menon et al. “Human and Pathogen Factors Associated with Chlamydia trachomatis-Related Infertility in Women”. In: *Clinical Microbiology Reviews* 28.4 (2015), pp. 969–985. URL: https://cmr.asm.org/content/28/4/969.

[18] Pippa Oakeshott et al. “Randomised controlled trial of screening for Chlamydia trachomatis to prevent pelvic inflammatory disease: the POPI (prevention of pelvic infection) trial”. In: *BMJ* 340 (2010). URL: https://www.bmj.com/content/340/bmj.c1642.

[19] L. R. Ranker, J. M. Petersen, and M. P. Fox. “Awareness of and potential for dependent error in the observational epidemiologic literature: A review”. In: *Ann Epidemiol* 36 (2019), 15–19.e2. ISSN: 1047-2797. DOI: 10.1016/j.annepidem.2019.06.001.

[20] Joanne Reekie et al. “Risk of Ectopic Pregnancy and Tubal Infertility Following Gonorrhea and Chlamydia Infections”. In: *Clinical Infectious Diseases* 69.9 (2019), pp. 1621–1623. ISSN: 1058-4838. DOI: 10.1093/cid/ciz145. URL: https://dx.doi.org/10.1093/cid/ciz145.

[21] Joanne Reekie et al. “Risk of Pelvic Inflammatory Disease in Relation to Chlamydia and Gonorrhea Testing, Repeat Testing, and Positivity: A Population-Based Cohort Study”. In: *Clinical Infectious Diseases* 66.3 (2018), pp. 437–443. ISSN: 1058-4838. DOI: 10.1093/cid/cix769.

[22] J. Riha et al. “Who is being tested by the English National Chlamydia Screening Programme? A comparison with national probability survey data”. In: *Sexually Transmitted Infections* 87.4 (2011), pp. 306–311. ISSN: 1368-4973. DOI: 10.1136/sti.2010.047027. URL: https://dx.doi.org/10.1136/sti.2010.047027.

[23] Leonard A. Stefanski and Raymond J. Carroll. “Covariate Measurement Error in Logistic Regression”. In: *Ann. Statist.* 13.4 (1985), pp. 1335–1351. ISSN:

0090-5364. DOI: 10.1214/aos/1176349741. URL: https://projecteuclid.org:443/euclid.aos/1176349741.

[24] Lars Westrom et al. “Pelvic Inflammatory Disease and Fertility”. In: *Sexually Transmitted Diseases* 19.4 (1992), pp. 185–192.

[25] D. P. Wilson. “Mathematical modelling of Chlamydia”. In: *Proc. of 11th Computational Techniques and Applications Conference CTAC-2003*. Ed. by Jagoda Crawford and A. J. Roberts. Vol. 45, pp. C201–C214.

[26] Grace Y. Yi et al. “Parametric Regression Analysis with Covariate Misclassification in Main Study/Validation Study Designs”. In: *The International Journal of Biostatistics* 15.1 (2019). ISSN: 1557-4679. DOI: 10.1515/ijb-2017-0002.

[27] Xiang Zhou and Jerome P. Reiter. “A Note on Bayesian Inference After Multiple Imputation”. In: *The American Statistician* 64.2 (2010), pp. 159–163. DOI: 10.1198/tast.2010.09109. eprint: https://doi.org/10.1198/tast.2010.09109. URL: https://doi.org/10.1198/tast.2010.09109.

4

Nonlinear modelling of Assisted Reproductive Technologies and testing on low birth weight pregnancy outcomes.

4.1 Introduction

In our previous chapters, the reproductive burden of testing for Chlamydia and/or Gonorrhea by using pregnancy substantiated with a birth record as a measure of fertility was discussed. This finding of increased reproductive burden was supported by the usage of low birth weight and preterm birth as further measures of reproductive health; women with a history of testing were more likely to have pregnancies with a gestational period of less than 37 weeks, and babies weighing less than 2500g.

These thresholds are recommended as measuring adverse perinatal outcomes [1], and the relationship between Chlamydia, Gonorrhea and these adverse perinatal outcomes have been described previously in the literature [10]. Adverse obstetric outcomes of preterm birth and small for gestational age births were measured for women with a history of testing for chlamydia before and during their pregnancy, for a cohort of 101,558 women from Western Australia. Women with a history of testing for chlamydia, prior to pregnancy, had increased odds of spontaneous preterm birth compared to women with a test for chlamydia during their

pregnancy when estimates were fully adjusted for age, area of residence, socioeconomic status, smoking during pregnancy, other infections and ethnicity. However, there was no clear distinction between women with a history of testing prior to pregnancy and women with no testing records for preterm birth. Women with a testing history prior to pregnancy had reduced odds of a small for gestational age birth, compared to women who were tested during pregnancy, but there was no distinction to women with no testing record. Overall, there was no difference in the odds of either outcome between women with negative and positive tests.

Preterm birth and low birth weight were more common for pregnancies conceived for IVF, and more common in couples with female-factor infertility compared to male-factor infertility, in an analysis of 18,429 Australian birth records from 1996 to 2000 [15]. The incidence of low birth weight was 2.6 times that of the general population for infants conceived with ART that were born 37 weeks or later [14]. A unique feature of our data is that it contains information on pregnancies that utilised Assisted Reproductive Technologies (ART), which has also been associated with risk factors for preterm birth and low birth weight pregnancies [16, 8, 5]. The underlying etiology of the perinatal outcomes for women accessing ART to give birth remains unknown. It is not clear if the use of the technology or the underlying infertility is the main cause [15, 6], although behavioural characteristics of the mother have been proposed to be more likely to determine low birth weight [9].

Previous chapters demonstrated a substantial reproductive burden of Chlamydia and Gonorrhea on pregnancy. Prior population-based data linkage studies that have shown an association between testing and PID have also been useful in determining the link between STI testing history and infertility. Whilst the Uppsala Women's cohort study showed a linkage between infertility (in the general sense of the term), the population study based on Denmark showed a linkage between chlamydia testing history and tubal factor infertility [7, 4]. Tubal factor infertility is thought to be a direct consequence of tubal scarring related to a pro-inflammatory response to the presence of a pathogen in the upper genital tract, which will be investigated in later chapters. A linkage between chlamydia and gonorrhea testing, in the Australian context, has also been shown [12].

This chapter investigates the impacts of testing, infection and ART usage for our cohort of pregnant women in QLD, Australia on the perinatal outcome of birth weight, conditional on gestational age. The data collected and examined in Chapters 2 and 3 contains key features of perinatal outcomes that allows for further examination the reproductive health burden of the two pathogens of interest. The analysis aims to investigate the impact of testing for Chlamydia and/or Gonorrhea on the birth weights of babies, once the impact of gestational weeks is accounted for through a mathematical model. We hypothesise that the impact of testing will be reflected in reduced birth weights, which can be considered as independent from the mediating effects of early pregnancies.

4.2 Methods

As described in the introduction, a low birth weight pregnancy is defined as any birth where the baby weight is < 2500g [1]. The use of a binary outcome measure allows for ease of interpretation, as a logistic regression model can produce an odds ratio that describes the change in odds of low birth weight for certain predictors, without having to interpret, for example, a mean change in birth weight produced from a linear regression. However, the discretisation of the outcome variable also wastes a significant amount of information that could otherwise inform our inferences.

In this chapter, we propose a model that describes continuous birth weight as conditional on a parametric nonlinear curve of gestational age of the pregnancy. The advantage of this method is that it fully utilises the available information. By estimating the model within a Bayesian framework, we show how we can easily interpret parameters relating to the continuous outcomes, and provide a posterior probability of low birth weight for ease of interpretation. We then extend the nonlinear model by including on the parameters of this curve, which we then use to address the scientific question posed in the introduction. In total, we show that modelling an outcome from first principles makes full use of the available information and can be used to answer multiple scientific questions simultaneously.

The data used is derived from the datasets constructed for use in Chapter 2, including identical methods for cohort construction, linkage methodologies and correction for errors and outcome definitions. Overall, birth weights from 65,481 pregnancies were modelled from women in QLD between 2005 and 2013.

The analysis in this chapter was completed using R version 3.6.3. Models were specified and run using the brms package (v2.13).

4.3 Results

4.3.1 Modelling gestational weeks and birth weight

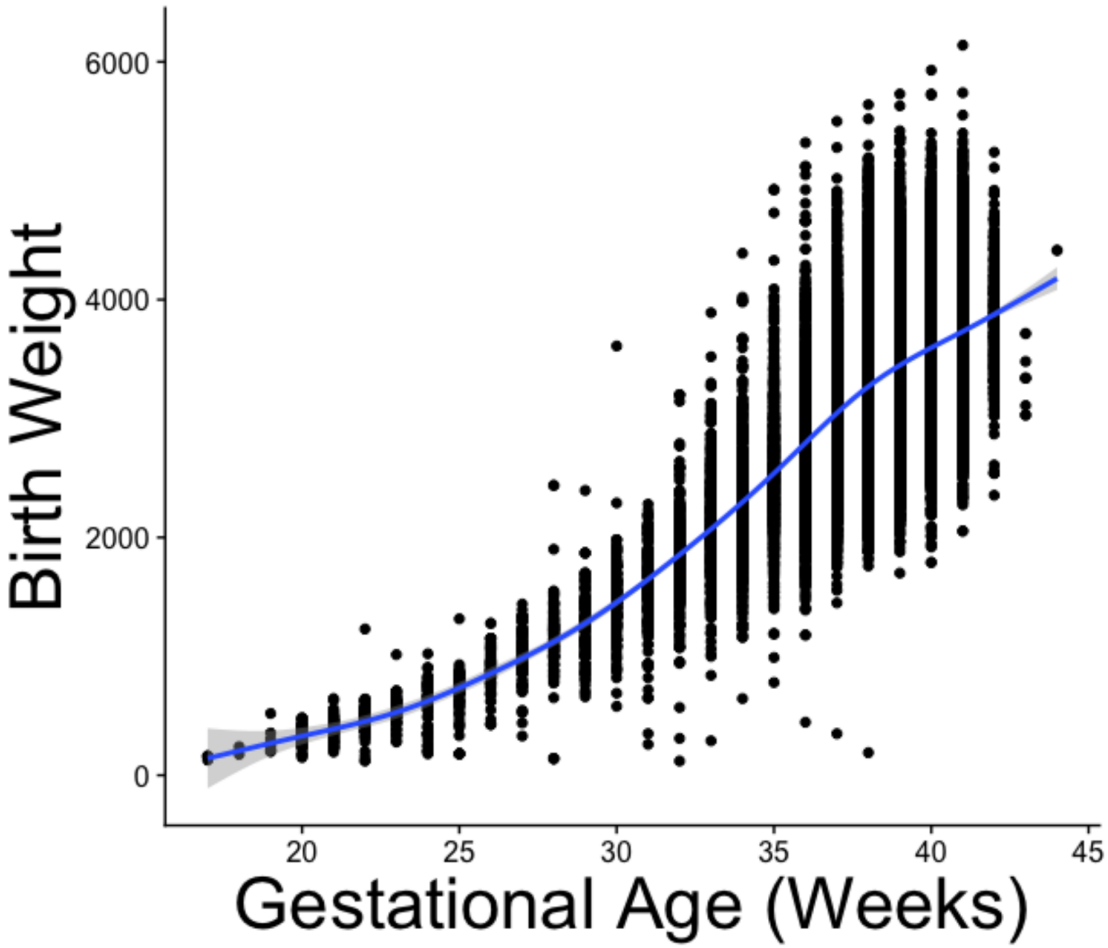

Figure 4.1: **The relationship between gestational age and birth weight.** The line plotted is derived from a smoothing spline.

The definition of low birth weight as any birth weight below 2500g is a

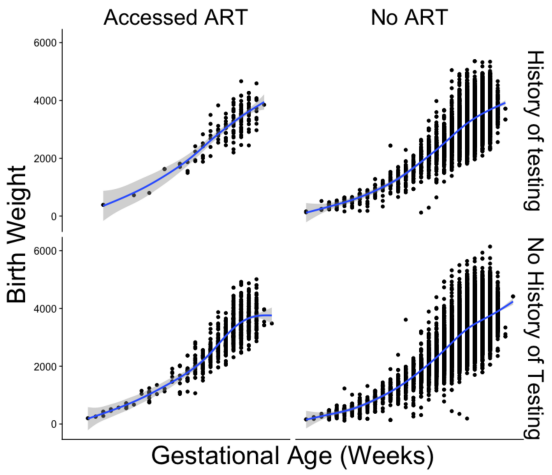

Figure 4.2: **The relationship between gestational age and birth weight.** The line plotted is derived from a smoothing spline. Compares women with no ART usage, and usage of ART, no history of testing and a history of testing.

standard one, but the discretisation of the outcome fails to utilise all of the available information. Making use of all the available information when constructing a model allows for a deeper understanding of the scientific problem, without wasting data.

The relationship between gestational age and birth weight is nonlinear, starting at a low value before having a period of greatest increase in the middle of the standard gestational period, before tapering close to full term (Figure 4.1). This relationship might also differ across different levels of the ART and testing factors of interest (Figure 4.2). For example, a pregnancy that utilised ART might be slightly lower in weight conditional on reaching full term, but otherwise be equivalent to natural births for shorter gestational periods.

To investigate the differing relationship between gestational age and birth weight for different pregnancy factors, we derive a mathematical model for birth weight on gestational age, and then run regressions on the parameters of the model. A mathematical model can be easily derived from some first principles about the process of gestation and birth weight. These principles are:

- There is a physical constraint that determines the minimum weight of baby (this means that a baby must have at least positive weight)

- There is another physical constraint that determines that maximum weight of baby when he/she is born (a baby cannot be 100kg, for example)
- The rate of growth of a baby in utero is higher later in the pregnancy compared to earlier, but then slows down again as the pregnancy reaches full term
- The period of greatest growth may differ from pregnancy to pregnancy
- The variation of baby weights between pregnancies will be larger as the gestational age approaches full term

An appropriate model for the phenomena described above is the four parameter logistic curve

$$f(x) = \beta_0 + \frac{\beta_1 - \beta_0}{1 + \exp(-\beta_3(x - \beta_2))},$$

where β_0 governs the lower asymptote of the curve (interpretable as the lowest possible birth weight when the gestational period is short), β_1 is the carrying capacity (or the largest possible birth weight when the gestational period is long), β_2 is the midpoint of the curve (when the period of greatest growth occurs) and β_3 is the growth rate of the curve (how quickly a baby will grow when the period of growth occurs).

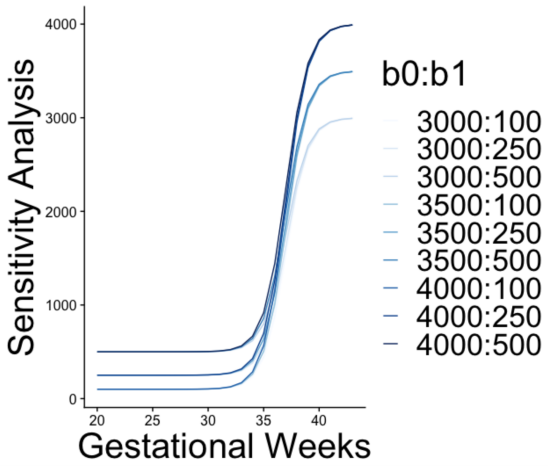

Figure 4.3: **The sensitivity of the model to the asymptote parameters.**

Before translation into a statistical model for fitting, a sensitivity analysis of the birth weight outcome to different parameter values of the mathematical model

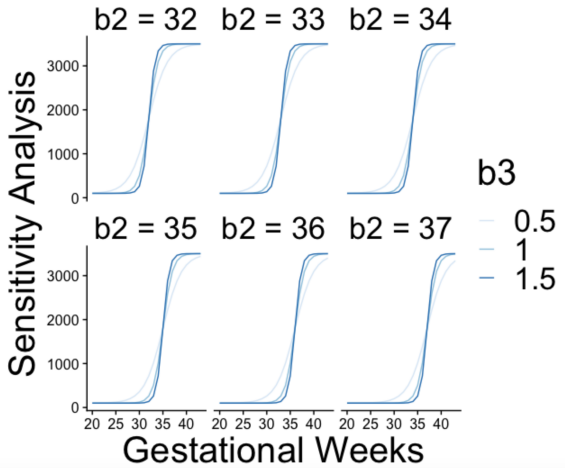

Figure 4.4: **The sensitivity of the model to the steepness and midpoint parameters.**

was undertaken. An increase in either of the two asymptote values (β_0 and β_1) would increase the outcome experienced at either end of the gestational age period (Figure 4.3). However, changes in either of these two parameters would not hugely impact the expected birth weights in the middle of the curve.

In contrast, the choice of midpoint and steepness parameters (β_2 and β_3) were highly influential in the expected birth weights for the middle gestational ages (Figure **??**). Importantly for the setup of the fitting procedure, much smaller changes in these parameters resulted in much larger changes in the birth weight outcome, compared to changes in β_0 and β_1. This observation would suggest to use prior distributions with a much tighter support compared to β_1. It would also suggest that a small change in these effects would be considered as more scientifically significant in comparison to larger changes in the asymptote parameters.

To translate this model into a full statistical model (for a single covariate of gestational age), we allow the curve $f(x)$ to describe the mean of the outcome distribution. Each parameter in the model must also be assigned appropriate priors. To address the fifth observation above, we introduce a linear predictor for the variance of the outcome distribution. The variance linear predictor has an intercept term and a slope term that refers to the increasing variance with increasing gestational age. The variance related terms must also be assigned priors. Combining these ideas, we

can express the curve in the following fashion

$$
\begin{aligned}
&birthweight_i \sim N(\mu_i, \sigma_i^2) \\
&\mu_i = \beta_0 + \frac{\beta_1 - \beta_0}{1 + \exp(-\beta_3(gestational\ age_i - \beta_2))} \\
&\beta_0 = \beta_{0,loc} + \beta_{0,int} \\
&\beta_1 = \beta_{1,loc} + \beta_{1,int} \\
&\beta_2 = \beta_{2,loc} + \beta_{2,int} \\
&\beta_3 = \beta_{3,int} \\
&log(\sigma_i^2) = \gamma_{int} + \sum_{j=1}^{J} \gamma_j gestational\ age_i \\
&\beta_{0,loc} = 105 \\
&\beta_{1,loc} = 3700 \\
&\beta_{2,loc} = 35 \\
&\beta_{0,int} \sim N(0, 0.5) \\
&\beta_{1,int} \sim N(0, 250) \\
&\beta_{2,int} \sim N(0, 0.5) \\
&\beta_{3,int} \sim N(0, 0.5) \\
&\gamma_{int} \sim N(0, 1) \\
&sd(\gamma)_j \sim N(0, 1)
\end{aligned}
\tag{4.1}
$$

The observed outcome is assumed to be gaussian (although other distributions may also be appropriate), conditional on the mean value μ_i that is derived from the nonlinear curve of gestational age $f(x)$.

The variance of the outcome distribution is predicted by a linear predictor and transformed to the appropriate non-negative support by using a log link as shown above. The intercept term and the standard deviation of the smoothing term is provided with standard normal $N(0, 1)$ priors, which are moderately informative on the log scale as we want to constrain the magnitude of the variance component of the regression. The relationship between variance and gestational age is modelled with a smoothing spline, to allow the relationship to be flexibly nonlinear. We do not give the variance part of the regression a pre-defined functional form as there is

an insufficient amount of principle information from which to derive such a form.

The parameter-level regression is decomposed into a constant term, and an intercept term. The constant term is assumed for each parameter to allow for identifiability of the model. Any modification of the parameters of the key part of the model will lead to un-identifiability of the model, if the parameter is allowed to vary over too large a domain. For example, a smaller value of β_3 represents a slower growth rate at the midpoint of the curve. Although this does not impact the carrying capacity, it may, if small enough, lower the value of the curve for $x = 40$, which represents a full term pregnancy (Figure 4.4). Extra care is needed in order to fit the model.

The constant term is broadly informed from other available data sources. Although the constant term is fixed in this model, in future the model could allow this term to be informed by the data as well, even if strong prior assumptions are made. The inclusion of the constant term also allows the priors for the intercept term to be centered on zero. Diagnosis of overly informative prior choice is therefore made possible. A posterior distribution that has a standard deviation close to the standard deviation of the intercept prior would suggest that the prior is poorly specified or overly informative. The intercept term for each parameter level regression is then given a prior with a variance that is informed by the relative scale of the parameter and a sensitivity analysis of the model.

Variable	Estimate	Est.Error
$\beta_{0,int}$	0.08	0.50
β_{1int}	849.17	19.05
$\beta_{2,int}$	-1.15	0.05
$\beta_{3,int}$	0.21	0.00
γ_{int}	6.11	0.00
$\gamma_{smooth\ of\ gestational\ age}$	1.64	0.93

Table 4.1: **Summary of the nonlinear model estimates.** No covariate terms are included in this model.

The estimated parameters differ slightly from the prior distributions (Ta-

ble 4.1, Figure 4.8). The posterior mean of $\beta_{0,int}$ was 0.08, which implies that the total modelled value of β_0 was 105.08 (95% CrI 104.1 - 106.3). Similarly, the posterior mean of β_2 was equal to 33.85 (95% CrI 33.75 - 33.95), and β_3 was equal to 0.21 (95% CrI 0.21 - 0.22). The mean posterior value for the upper asymptote (β_1 was 4549.17 (95% CrI 4510.85 - 4586.3)).

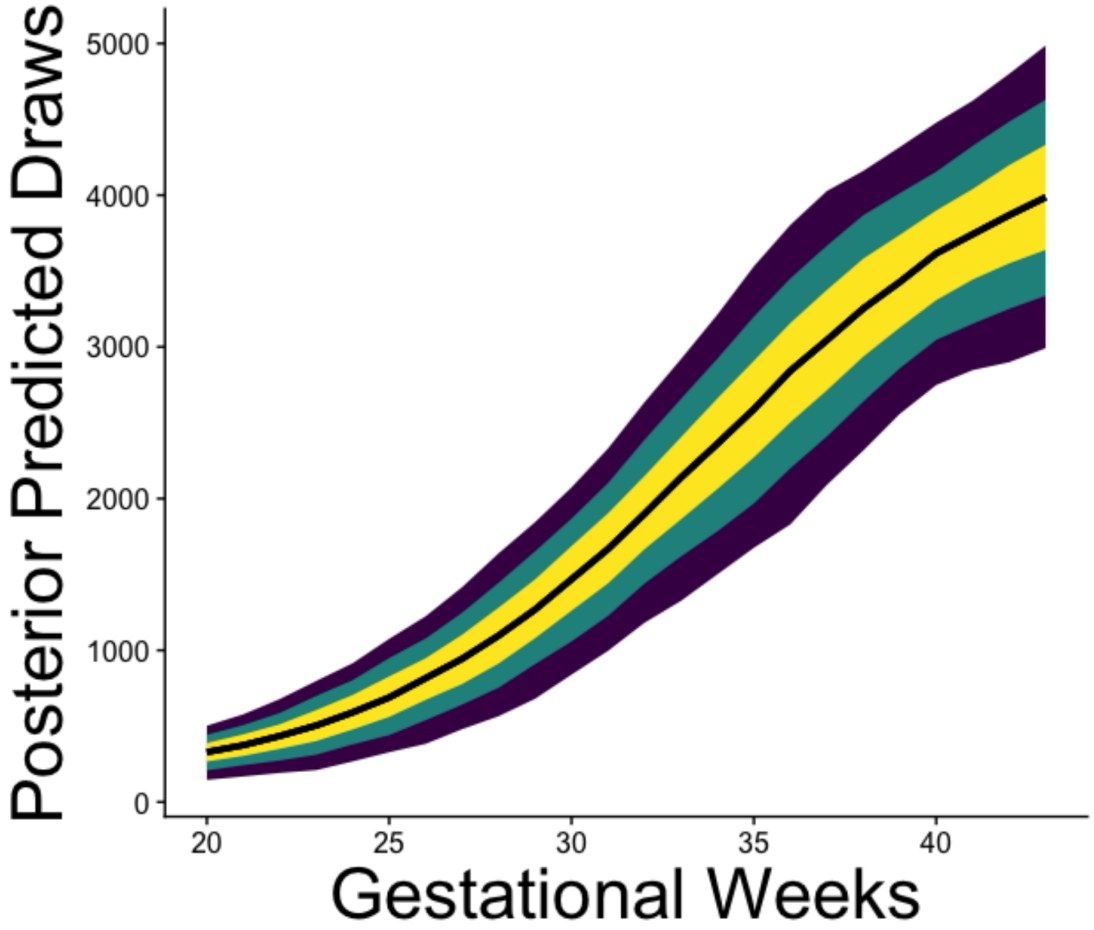

Figure 4.5: **Draws from the posterior predictive distribution of the relationship between gestational age and birth weight.**

Taking draws from the posterior predictive distribution, we can then view the modelled relationship between gestational age and birth weight (Figure 4.5). The overall shape of the curve closely matches the observed data in Figure 4.1, longer gestational ages imply higher birth weights on average, and the variance grows as the gestational period becomes larger (Figure 4.6). The curve does not plateau (that is, reach is upper asymptote) at full-term pregnancies (40 weeks).

To investigate the ability of the model formulation to capture the overall

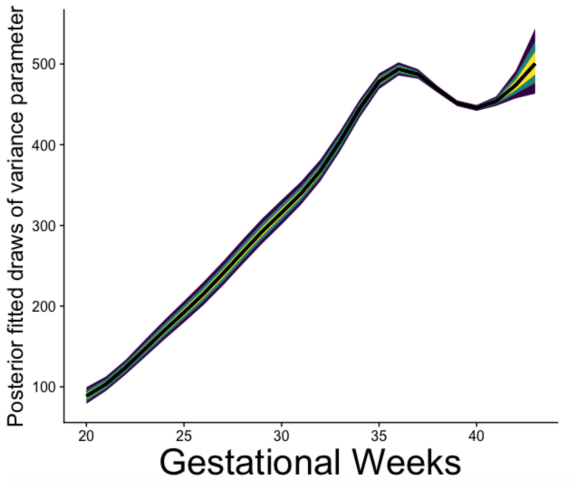

Figure 4.6: **Draws from the posterior predictive distribution of the relationship between gestational age and the variance birth weight.**

structure of the gestational age and birth weight process, we examined the model fit by comparing the observed distribution of birth weights in the data to repeated draws from the posterior distribution. The model appears to capture well the left tail of the distribution, and approximates the shape of rest of the distribution well (Figure 4.7(A)). The mean of the distribution is matched by the mean of the posterior distribution (Figure 4.7(B)).

The minimum and the maximum of the distribution is understated in the posterior distribution compared to the observed data (Figure 4.7(C), Figure 4.7(D)). In particular, the minimum of the posterior distribution is sometimes negative, which implies impossible birth weights below zero. This is a consequence of the choice of the outcome distribution. A more appropriate choice might be to model birth weights on the log scale, and have the nonlinear curve operate on the expected value of the distribution multiplicatively.

We would expect that a large portion of the simulated minimums from the dataset would appear below the observed minimum. Similar behaviour would be expected for the distribution of the maximums, however a majority of the simulated maximums are below the observed one. The implication is that the model does not yet appropriately capture the full range of birth weights effectively, and that improvements to the model setup would capture the observed maximum within the

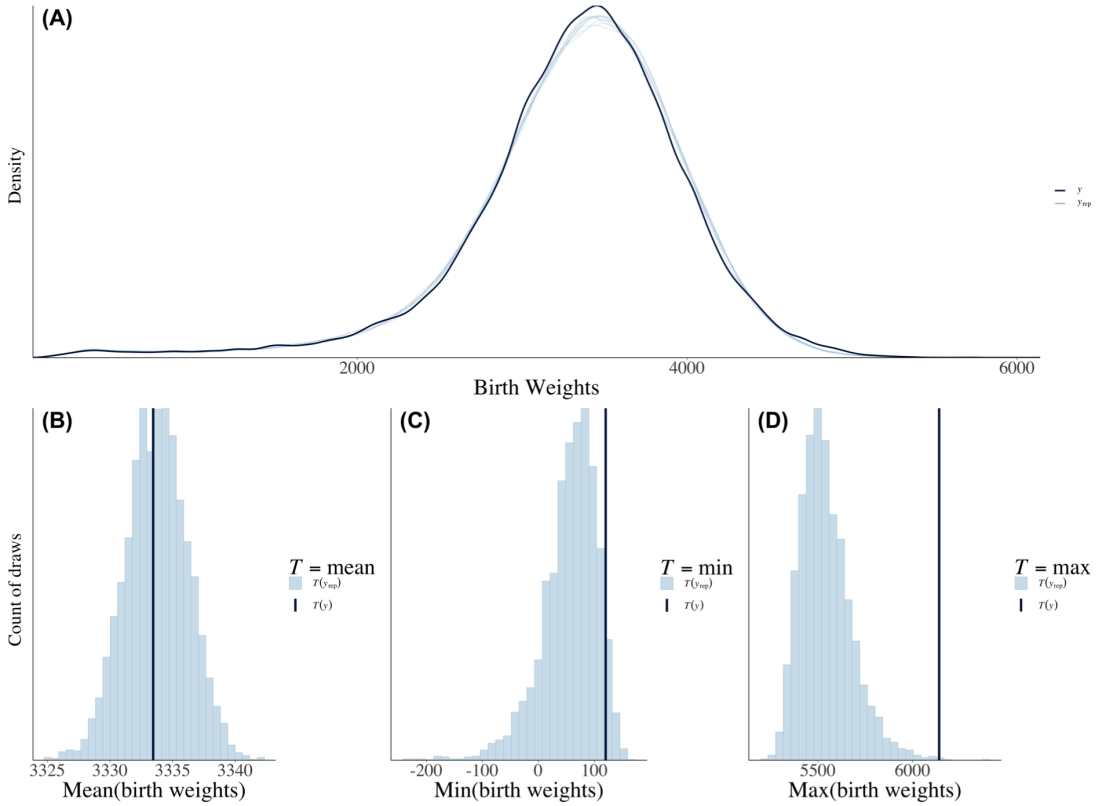

Figure 4.7: **A comparison of the observed distribution of birth weights and draws from the posterior distribution.** Draws are represented in blue, 10 replications from the posterior are shown in this figure. (A) Shows the full distribution. (B) Shows the distribution of the mean against the observed mean (black line). (C) Shows the distribution of the minimum. (D) Shows the distribution of the maximum.

distribution of simulations from the model.

4.3.2 Incorporating covariates into the model structure

As noted within the introduction, we are interested in comparing the overall effects of ART usage and a history of testing on birth weights. We extend the model developed in the previous subsection, by introducing these covariates into the linear predictors of each of the parameters of the overall nonlinear model. Each parameter in the model $(\beta_0, \beta_1, \beta_2, \beta_3)$, is allowed to vary as a function of ART usage and testing history. The parameter values for each group then determines the expected birth weight for a particular gestational age. As with the previous model, the variance

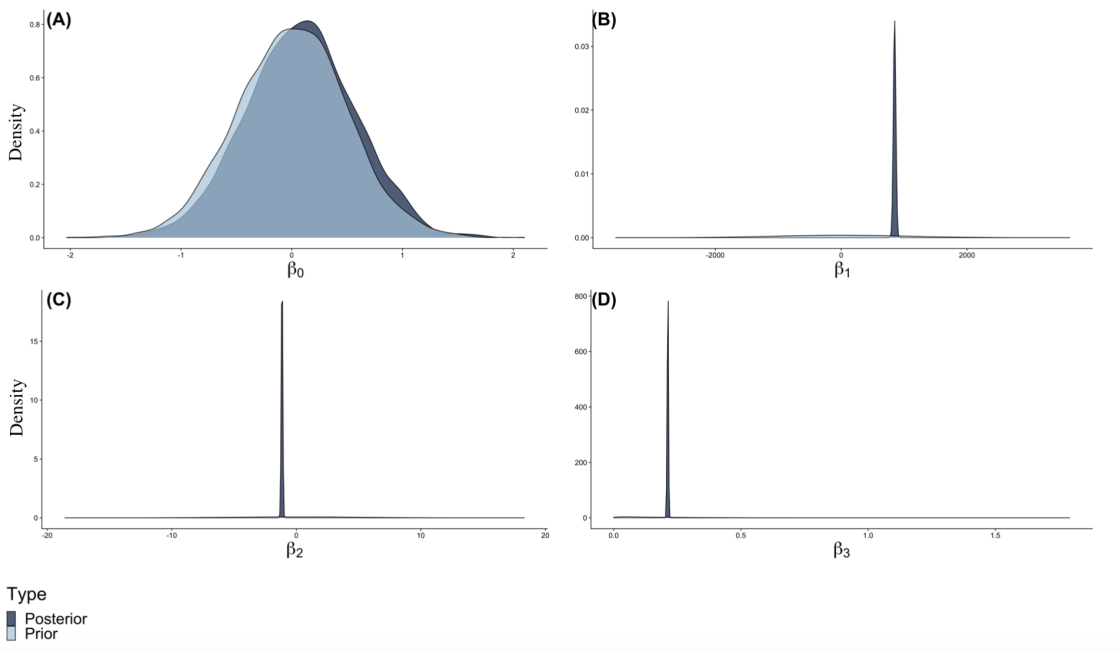

Figure 4.8: **A comparison of the prior and posterior distributions for each parameter of the nonlinear model.** (A) β_0 (B) β_1 (C) β_2 (D) β_3

is allowed to flexibly vary as a function of gestational age by using a spline for the linear predictor of the log of the variance term.

To translate this concept into a full statistical model (for the covariate of gestational age, and additional covariates ART_i and $testing_i$), we can express the

curve in the following fashion

$$
\begin{aligned}
&birthweight_i \sim N(\mu, \sigma^2) \\
&\mu = \beta_0 + \frac{\beta_1 - \beta_0}{1 + \exp(-\beta_3(gestationalage_i - \beta_2))} \\
&\beta_0 = \beta_{0,loc} + \beta_{0,int} + \beta_{0,ART} ART_i + \beta_{0,testing} testing_i \\
&\beta_1 = \beta_{1,loc} + \beta_{1,int} + \beta_{1,ART} ART_i + \beta_{1,testing} testing_i \\
&\beta_2 = \beta_{2,loc} + \beta_{2,int} + \beta_{2,ART} ART_i + \beta_{2,testing} testing_i \\
&\beta_3 = \beta_{3,int} + \beta_{3,ART} ART_i + \beta_{3,testing} testing_i \\
&log(\sigma^2) = \gamma_{int} + \sum_{j=1}^{J} \gamma_j gestational\ age_i \\
&\beta_{0,loc} = 105 \\
&\beta_{1,loc} = 3700 \\
&\beta_{2,loc} = 35 \\
&\beta_{0,int}, \beta_{0,ART}, \beta_{0,testing} \sim N(0, 0.5) \\
&\beta_{1,int}, \beta_{1,ART}, \beta_{1,testing} \sim N(0, 250) \\
&\beta_{2,int}, \beta_{2,ART}, \beta_{2,testing} \sim N(0, 0.5) \\
&\beta_{3,int}, \beta_{3,ART}, \beta_{3,testing} \sim N(0, 0.5) \\
&\gamma \sim N(0, 1) \\
&sd(\gamma)_j\ gestational\ age_i \sim N(0, 1)
\end{aligned}
\tag{4.2}
$$

Each parameter in the nonlinear curve can be regressed as a function of observed covariates. The parameter-level regression is decomposed into a constant term, an intercept term and a covariate term. The constant term is assumed for each parameter to allow for identifiability of the model, particularly the parameters β_2 and β_3. The constant term is broadly informed from other available data sources. Although the constant term is fixed in this model, in future the model could allow this term to be informed by the data as well, even if strong prior assumptions are made. The intercept term for each parameter level regression is then given a prior that is cantered on zero, with a variance that is informed by the relative scale of the parameter and a sensitivity analysis of the model. The covariate term (or terms in expansions of the model) is given the same prior for each parameter. As shown in

Figures 4.3 4.4, modifying an individual parameter at a time modifies other observed characteristics of the curve, so extra care is needed in order to fit the model. For example, a smaller value of β_3 represents a slower growth rate at the midpoint of the curve. Although this does not impact the carrying capacity, it may, if small enough, lower the value of the curve for $x = 40$, which represents a full term pregnancy (Figure 4.4).

After confirming the model formulation appropriately captures the distribution of the outcome, the model was ran with covariates of ART usage, history of testing and the interaction of the two was included. Priors for the main effects were as per the formulation in equation 4.2.

4.3.3 Impacts of ART usage and testing history on birth weights

Variable	Estimate	Est.Error	95% CrI
$\beta_{0,int}$	0.07	0.52	-0.94, 1.06
$\beta_{0,ART}$	0.00	0.49	-0.97, 0.98
$\beta_{0,testing}$	0.02	0.51	-0.98, 1.00
$\beta_{1,int}$	833.71	20.57	793.30, 873.30
$\beta_{1,ART}$	147.85	95.58	-24.20, 345.54
$\beta_{1,testing}$	11.04	45.33	-75.68, 101.36
$\beta_{2,int}$	-1.22	0.06	-1.33, -1.11
$\beta_{2,ART}$	0.42	0.25	-0.03, 0.92
$\beta_{2,testing}$	0.16	0.13	-0.09, 0.41
$\beta_{3,int}$	0.22	0.00	0.21, 0.22
$\beta_{3,ART}$	0.00	0.01	-0.01, 0.01
$\beta_{3,testing}$	-0.01	0.00	-0.02, -0.01
σ^2_{int}	6.11	0.00	6.10, 6.11
$\sigma^2_{smooth\ of\ gestational\ age}$	1.65	0.96	-0.32, 3.54

Table 4.2: **Summary of the nonlinear model estimates.** ART usage and history of testing terms are included in this model.

Group	β_0 (95% CrI)	β_1 (95% CrI)	β_2 (95% CrI)	β_3 (95% CrI)
No ART usage, no history of testing	105.0 (104.0 - 106.0)	4534 (4493 - 4573)	33.8 (33.7 - 33.9)	0.216 (0.213 - 0.219)
ART usage, no history of testing	105.0 (104.0 - 106.0)	4679 (4507 - 4879)	34.2 (33.7 - 34.7)	0.216 (0.203 - 0.229)
No ART usage, history of testing	105.0 (104.0 - 106.0)	4544 (4467 - 4627)	33.9 (33.7 - 34.2)	0.204 (0.199 - 0.210)
ART usage, history of testing	105.0 (103.0 - 107.0)	4691 (4497 - 4902)	34.4 (33.9 - 34.9)	0.204 (0.190 - 0.219)

Table 4.3: **Summary of the nonlinear model parameters.** Each combination of ART usage and history of testing terms are shown, along with posterior mean and 95% posterior interval.

The summary of the estimated values for the model is in Table 4.2, and the estimated parameters for each group is summarised in Table 4.3. Overall, there was no huge distinction between groups for the value of β_0 (the lower asymptote). The value for β_1 was larger overall for women with ART usage. Women with a history of testing and/or ART usage had larger values of β_2, suggesting that the week of gestation where the period of greatest growth in baby weight occurred was later for these women. There was no difference in β_3 for ART usage, but women with a history of testing had a lower value for this parameter, suggesting that the rate of growth was smaller for women with a history of testing.

All parameters in the model had 95% credible intervals that covered 0 (with the exception of β_2 for women with a history of testing). However, some of the other model parameters had significant posterior mass that was greater (or less than) 0, which can be easily represented by computing the one-sided hypothesis using samples of the posterior distribution (Figure 4.9). As discussed above, there was no difference for β_0 for women with a history of testing or with ART usage. There was no noticeable difference for β_1 for women with a history of testing, but there was a 0.95 chance of β_1 being greater than 0 for women with ART usage. For β_2, there was a probability of 0.9 of the parameter being greater than 0 for women with a history of testing, and 0.97 for women with a ART usage. All of the samples of posterior distribution was below 0 for β_3 for women with a history of testing (4000 samples were used in total), but there was no shift in the posterior distribution of β_3 for ART usage.

The effects of ART usage and testing history can also be summarised graph-

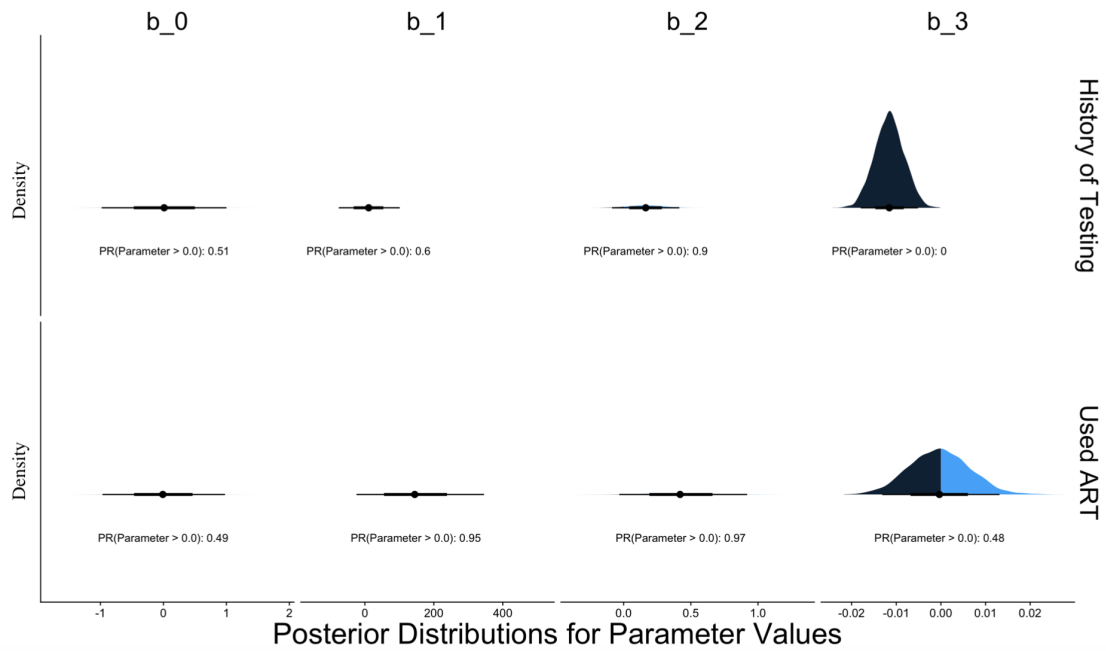

Figure 4.9: **Summary of the posterior distributions of the parameters from the nonlinear model with covariates.** Text shows the posterior probability of observing the effect being greater than 0.0

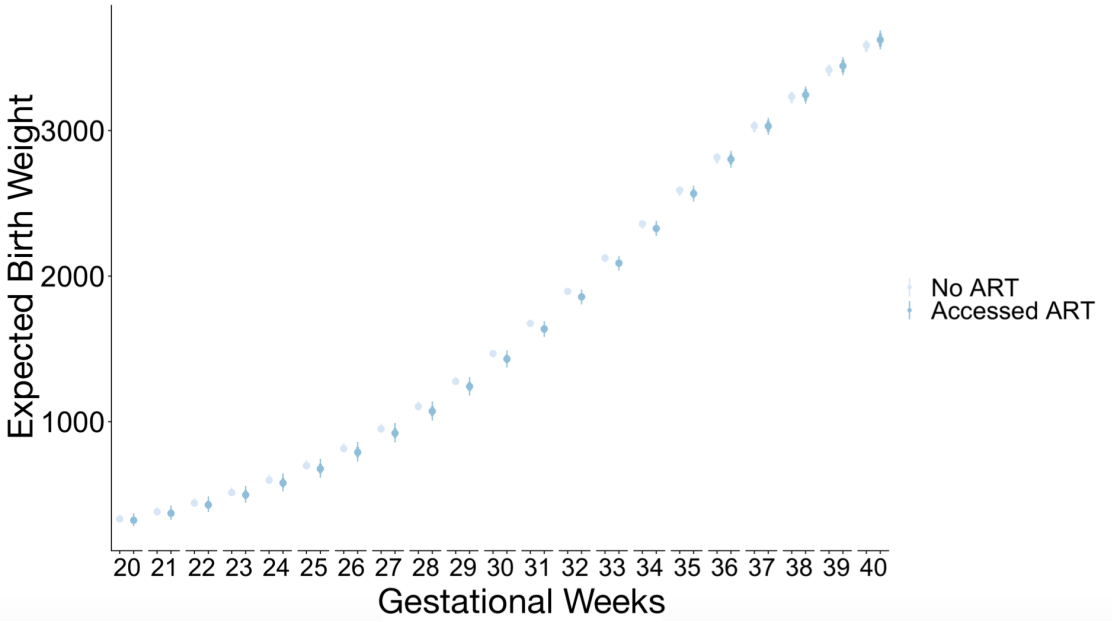

Figure 4.10: **Summary of the posterior distributions of the mean of birth weight across gestational age.** Compares pregnancies with and without ART usage.

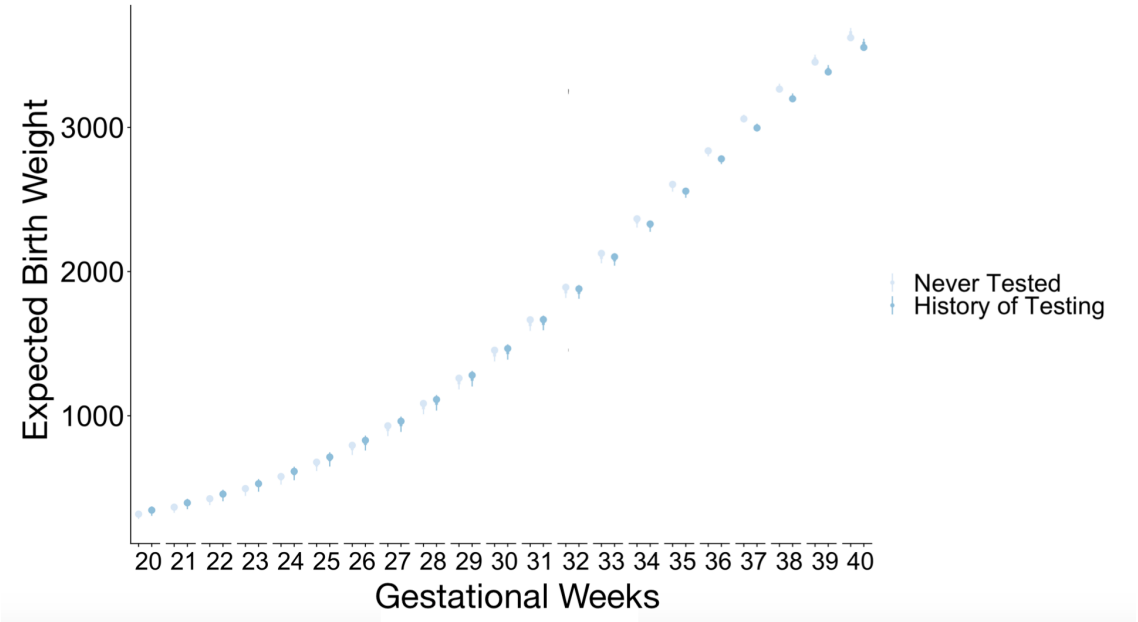

Figure 4.11: **Summary of the posterior distributions of the mean of birth weight across gestational age.** Compares pregnancies with and without a history of testing.

ically. Women with ART pregnancies are more likely to have lower birth weight babies than women with no ART usage, when the baby is born earlier in the gestational period (Figure 4.10), and more likely to have similar or higher birth weight babies when the pregnancy reaches full term. This is likely influenced by the higher values of β_1 and β_2 for this group (Figure 4.9, Table 4.3).

Women with a history of testing experience the converse. When a baby is born earlier in the gestational period they are more likely to weigh more when born to a woman with a history of testing, compared to women with no history of testing (Figure 4.11). This is influenced by the higher values of β_2, and lower values of β_3 for this group (Figure 4.9.

Although we don't observe 95% CI of parameters of the model that are fully greater or less than 0, we still see some evidence of a difference in birth weight conditional on gestational age for both ART usage and history of testing variables. The posterior probability of pregnancies with ART usage having lower birth weight than pregnancies without is high between weeks 20 and 36 of gestation (Table 4.4).

Week of gestation	Posterior probability ART < No ART	Posterior probability - Test
20	0.73475	0.00025
21	0.75075	0.00025
22	0.76625	0.00025
23	0.78000	0.00025
24	0.80150	0.00025
25	0.82550	0.00075
26	0.84650	0.00225
27	0.86775	0.00575
28	0.89275	0.01725
29	0.91025	0.05000
30	0.92625	0.15200
31	0.94025	0.41900
32	0.93975	0.80300
33	0.93125	0.98000
34	0.90775	0.99925
35	0.84800	1.0
36	0.70725	1.0
37	0.44200	1.0
38	0.15275	1.0
39	0.02200	1.0
40	0.00650	1.0

Table 4.4: **Posterior probability of the birth weight being less than the reference level of (respectively) no ART usage and no history of testing.**

The posterior probability of pregnancies with a history of testing having lower birth weight than pregnancies without is higher in contrast for pregnancies beyond 32 weeks until full term gestation.

4.4 Discussion

In this chapter, we have examined the overall impact of ART usage and a history of testing for Chlamydia and/or Gonorrhea on the birth weights of babies born to women in these groups. Although previous studies have assessed the impact of these two variables independently [16, 14, 8, 15, 5, 10, 9], this is the first time the interaction between the two has been examined. We have also introduced a mathematical model to study the overall distribution of birth weights given gestational age, that makes full use of the data without discretisation.

Women with a history of testing had overall lower birth weights when a pregnancy was carried to full gestational term, however there was no observed additional impact on birth weights due to testing history for lower gestational ages. It has been previously proposed that the impact of Chlamydia and Gonorrhea on fertility is mediated by tubal scarring [2, 3, 11, 13], and in Chapters 2 and 3 we determined that a history of testing was the most substantiated measure of exposure to infection, when using pregnancy as a measure of reproductive capability.

Women with ART usage generally do not have lower birth weight pregnancies when the pregnancy goes to full term. However when the baby is born before 37 weeks, there is moderate probabilities (less than 95% but greater than 50%) that the birth weight is likely to be lower than for similar pre-term births with no usage of ART. This may be an influence of the technology itself, or may be a result of the underlying fertility issues that preceded the need for those women accessing ART. As the average birth weight is lower for babies born to shorter gestational age pregnancies, this difference, although small, may be a greater critical risk to the ongoing health of the child.

Overall, the curves of birth weight and gestational weeks have slightly different characteristics between women with ART usage and a history of testing.

When interpreted in the context of previous studies of birth weights, ART and testing, the results suggest that the underlying demographic factors that impact overall fertility challenges faced by women in these groups are different. A recent analysis of babies born to women in Utah proposes that knowledge of women's behavioural characteristics are more likely to explain low birth weights than usage of medically assisted reproduction types, although their analysis shows that women who used ART are still likely to have lower birth weights after adjusting for key behavioural confounders [9]. Our analysis did not have access to these key variables of interest, and so the weakness of the evidence presented may be attributable to this fact. On the other hand, the above study showed the women accessing ART also had lower gestational ages, but did not account for this key fact when assessing birth weights. Other studies also fail to consider the two variables in a single model. It is possible that our results are attenuated because the model derived accounts for the relationship between gestational age and birth weight. The difference in birth weights between groups might be a result of gestational age. The lack of clarity here demonstrates the critical importance of using an appropriate model to analyse the underlying process. The model derived in this chapter should be applied to a number of distinct datasets in order to gain a better appraisal of the effects of infections, ART usage and other medically assisted reproductive treatments on birth weights.

In this chapter, we have derived a simple mathematical model of a key process that underlies our observed data. We have then turned this into a full statistical model (with the introduction of prior distributions and a likelihood function), and estimated key parameters of the model as functions of our key variables of interest. This approach has allowed us to interpret parameters on a scale that relates to the process under examination (for example, an increase in β_2 leads to a later period of growth in the pregnancy), and has allowed us to make full use of the data without wasting information due to unnecessary discretisation. This approach has major benefits in providing ease of interpretation of the results (especially compared to a non parametric nonlinear fit such as splines or gaussian processes), and allows for aggregation of prior knowledge into a mechanistic mathematical model.

There are some significant drawbacks to this approach that should also be noted. The choice of functional form for the mathematical model is not unique, and is difficult to identify from the data without additional work on appropriate model comparison. Similarly, the outcome distribution (which was assumed to be gaussian in this chapter) may be better represented by other choices (such as a Student-t for fatter tails, or a Gamma distribution for strictly non-negative values). These choices are not unique to this approach (in fact most modelling problems pose similar challenges), however the interaction between the two makes the space of possibilities wide to examine with principled reasoning on the model formulation. For example, if a Gamma distribution was chosen as the outcome, a three-parameter logistic curve might be the best choice of mathematical model, but a five-parameter curve might suit a Student-T distribution. The space of model choices is further expanded when variable selection is also considered. As with a typical regression model, this approach requires the selection and specification of variables, and which of these should be fixed to population-level and which should vary in a hierarchical manner. In order to fit this model, strong priors were required to ensure different parameters were identifiable, and so the feasibility of this modelling approach relies upon the presence of rich prior information.

4.5 Conclusion

In this chapter, we have used a statistical model to examine the relationship between ART usage and a history of testing for Chlamydia and/or Gonorrhea on the birth weights of pregnancies. The form of the statistical model followed a mathematical model which was in turn motivated by an examination of the process under consideration. Using this technique, we found that women with ART usage and a history of testing both were more likely to experience pregnancies with an impact on their birth weight, but that the impact depended on the length of the gestational period.

References

[1] Standard. 2006. URL: https://www.who.int/childgrowth/standards/technical_report/en/.

[2] B. Davies, K. Turner, and H. Ward. "Risk of pelvic inflammatory disease after Chlamydia infection in a prospective cohort of sex workers". In: *Sex Transm Dis* 40.3 (2013), pp. 230–4. ISSN: 0148-5717. DOI: 10.1097/OLQ.0b013e31827b9d75.

[3] B. Davies et al. "Heterogeneity in risk of pelvic inflammatory diseases after chlamydia infection: a population-based study in Manitoba, Canada". In: *J Infect Dis* 210 Suppl 2.Suppl 2 (2014), S549–55. ISSN: 0022-1899 (Print) 0022-1899. DOI: 10.1093/infdis/jiu483.

[4] Bethan Davies et al. "Risk of reproductive complications following chlamydia testing: a population-based retrospective cohort study in Denmark". In: *The Lancet Infectious Diseases* 16.9 (2016), pp. 1057–1064. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(16)30092-5. URL: https://doi.org/10.1016/S1473-3099(16)30092-5.

[5] Eugene Declercq et al. "Perinatal outcomes associated with assisted reproductive technology: the Massachusetts Outcomes Study of Assisted Reproductive Technologies (MOSART)". In: *Fertility and Sterility* 103.4 (2015), pp. 888–895. ISSN: 0015-0282. DOI: 10.1016/j.fertnstert.2014.12.119. URL: https://dx.doi.org/10.1016/j.fertnstert.2014.12.119.

[6] Laxmi A. Kondapalli and Alfredo Perales-Puchalt. "Low birth weight: is it related to assisted reproductive technology or underlying infertility?" In: *Fertility and Sterility* 99.2 (2013), pp. 303–310. ISSN: 0015-0282. DOI: 10.1016/j.fertnstert.2012.12.035. URL: https://dx.doi.org/10.1016/j.fertnstert.2012.12.035.

[7] N. Low et al. "Incidence of severe reproductive tract complications associated with diagnosed genital chlamydial infection: the Uppsala Women's Cohort Study". In: *Sex Transm Infect* 82.3 (2006), pp. 212–218.

[8] Akira Nakashima et al. “Implications of assisted reproductive technologies on term singleton birth weight: an analysis of 25,777 children in the national assisted reproduction registry of Japan”. In: *Fertility and Sterility* 99.2 (2013), pp. 450–455. ISSN: 0015-0282. DOI: 10.1016/j.fertnstert.2012.09.027.

[9] Alina Pelikh et al. “Medically Assisted Reproduction Treatment Types and Birth Outcomes”. In: *Obstetrics and Gynecology* 139.2 (2022), pp. 211–222. URL: doi:%2010.1097/AOG.0000000000004655.

[10] Joanne Reekie et al. “Chlamydia trachomatis and the risk of spontaneous preterm birth, babies who are born small for gestational age, and stillbirth: a population-based cohort study”. In: *The Lancet Infectious Diseases* 18.4 (2018), pp. 452–460. ISSN: 1473-3099. DOI: 10.1016/s1473-3099(18)30045-8.

[11] Joanne Reekie et al. “Hospitalisations for Pelvic Inflammatory Disease Temporally Related to a Diagnosis of Chlamydia or Gonorrhoea: A Retrospective Cohort Study”. In: *PLoS ONE* 9.4 (2014), e94361. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0094361.

[12] Joanne Reekie et al. “Risk of Ectopic Pregnancy and Tubal Infertility Following Gonorrhea and Chlamydia Infections”. In: *Clinical Infectious Diseases* 69.9 (2019), pp. 1621–1623. ISSN: 1058-4838. DOI: 10.1093/cid/ciz145. URL: https://dx.doi.org/10.1093/cid/ciz145.

[13] Joanne Reekie et al. “Risk of Pelvic Inflammatory Disease in Relation to Chlamydia and Gonorrhea Testing, Repeat Testing, and Positivity: A Population-Based Cohort Study”. In: *Clinical Infectious Diseases* 66.3 (2018), pp. 437–443. ISSN: 1058-4838. DOI: 10.1093/cid/cix769.

[14] Laura A. Schieve et al. “Low and Very Low Birth Weight in Infants Conceived with Use of Assisted Reproductive Technology”. In: *New England Journal of Medicine* 346.10 (2002), pp. 731–737. ISSN: 0028-4793. DOI: 10.1056/nejmoa010806.

[15] Yueping Alex Wang et al. “Preterm birth and low birth weight after assisted reproductive technology-related pregnancy in Australia between 1996 and 2000”.

In: *Fertility and Sterility* 83.6 (2005), pp. 1650–1658. ISSN: 0015-0282. DOI: 10.1016/j.fertnstert.2004.12.033.

[16] Xu K Xu et al. "Risk factors associated with preterm birth among singletons following assisted reproductive technology in Australia 2007–2009–a population-based retrospective study". In: *BMC Pregnancy and Childbirth* 14.1 (2014). ISSN: 1471-2393. DOI: 10.1186/s12884-014-0406-y. URL: https://dx.doi.org/10.1186/s12884-014-0406-y.

5

Within host modelling of Chlamydia

5.1 Introduction

Chlamydia trachomatis is a bacterial pathogen and the cause of the most common notifiable sexually transmitted infection. There is considerable evidence to indicate that a Chlamydial infection in women may cause Pelvic Inflammatory Disease (PID), and that these women are more likely to experience tubal factor infertility [4]. The development of tubal pathology is hypothesised to be a result of a Chlamydial infection ascending beyond the cervix and into the upper reproductive tract, combined with a pro-inflammatory response to the pathogen [7].

The low rate of pathology development in women with infections suggests that a number of host response and pathogen factors moderates ascension. The immune system response is particularly critical in determining the development of pathology [6].

An infection develops by extracellular Elementary Bodies (EBs) entering into a cell. Inclusions may have many bacteria per cell. Once within the host cell, the metabolically inert EB differentiates into the Reticulate Body (RB) form of Chlamydia. The RBs replicate by binary fission for some cycles of division, until some RBs convert back into EBs while others begin to replicate. The cycle of development continues until approximately 40–48 hours post infection, after which EBs continually exit the cell in an extrusion, which leaves the host cell intact, or via the lytic cycle

in which the host cell releases a burst of EBs. The burst of EBs released can then go on to infect new cells, thus beginning the cycle again [1].

The micro-population dynamics of a bacterial infection and the associated host cell response have been represented by a system of ordinary differential equations [12]. More realistic behaviour was represented with a spatial dimension (represented with a set of partial differential equations) [5]. Host response scenarios can be modelled by changing parameter values of particular equations, which in turn yields insights into the long-term behaviour of the system under different conditions. However, this modelling approach does not capture the randomness of the initial phase of an infection.

As a precursor to explaining PID and infertility, we must describe the requirements for an infection to become established and to survive in the mucosal layer of the cervix for a sufficient period of time until ascension can occur and pathology can develop. In the model described in this chapter, we explicitly keep track of the count of infected cells and account for the stochastic elements present in the development of an infection. An understanding of the dynamics of the process, as described by simulations from model, will yield an insight into certain aspects of the disease process.

5.2 Modelling Chlamydial Dynamics

The most basic theoretical framework of the host-pathogen response for chlamydial infections considers three interacting populations of the bacterium, the host target cell and the host immune response. This system is an example of the predator-prey relationships common in ecology, most notably described by the Lotka-Volterra equations.

The concept was adapted to viral infections by Nowak and May [8], in which the presence of a virus increased an immune response, which in turn controlled the pathogen and so on. This extension of predator-prey systems for viral dynamics was considered for various functional forms of relationships, and for systems containing other populations such as the production of distinct immune responses.

Analysis of models of predator-prey systems (described by a system of Ordinary Differential Equations) can be used to determine the relationships between parameters and properties of the systems, most commonly in the reproductive ratio R_0 which gives the combination of parameters needed for a population to grow to a sustainable level. Equilibria of the system can also be characterised by their stability.

Chlamydia is an intracellular obligate parasite, so many of its features are shared with viruses despite being a bacterium, such as reproduction occurring only within a target cell. Therefore, many of the systems used to describe viral dynamics can be adapted to Chlamydia.

The micro-population dynamics of a bacterial infection and the associated host cell response have been represented by a system of ordinary differential equations [12]. We describe the basic model as the dynamics of three populations of cells, $E(t)$ the number of uninfected target cells, $I(t)$ the number of infected cells and $C(t)$ the number of Chlamydial Elementary bodies.

New infective bodies are produced as a proportion of those currently residing in infected cells. At the same time, some naturally die, and others infect new uninfected cells. Uninfected target cells have a natural birth and death process. Uninfected target cells can also become infected by Chlamydial bodies. Infected target cells are created by the infection process, and can die via the burst process are the completion of the chlamydial development cycle or destruction from the host immune response.

$$\begin{aligned}
\frac{dC}{dt} &= P\kappa_2 I(t) - \mu C(t) - \kappa_1 C(t)E(t), \\
\frac{dE}{dt} &= P_E - \delta_E E(t) - \kappa_1 C(t)E(t), \\
\frac{dI}{dt} &= \kappa_1 C(t)E(t) - \gamma I(t) - \kappa_2 I(t).
\end{aligned}$$

In this model, P represents the number of chlamydial particles released from infected cells at a rate of κ_2 and κ_1 is the rate of epithelial cell infection.

Epithelial cells are produced at a rate P_E and δ_E is the rate of epithelial cell natural death. We let μ be the clearance rate by macrophages, and γ be the rate of clearance of infected cells due to cell-mediated immunity.

As per the original publication, this model has two steady states. The first is trivial, representing the situation the infection has been eradicated by the host $C(t) = 0, I(t) = 0, E(t) = \frac{P_E}{\delta_E}$. The other steady state is non-trivial

$$\begin{aligned} C(t) &= \frac{P_E[(P-1)\kappa_2 - \gamma]}{\mu(\gamma + \kappa_2)} - \frac{\delta_E}{\kappa_1}, \\ E(T) &= \frac{\mu(\gamma + \kappa_2)}{\kappa_1[(P-1)\kappa_2 - \gamma]}, \\ I(t) &= \frac{P_E}{\gamma + \kappa_2} - \frac{\delta_E \mu}{\kappa_1[(P-1)\kappa_2 - \gamma]}. \end{aligned}$$

The R_0 for this model can be expressed as

$$R_0 = \frac{P}{(1 + \gamma/\kappa_2)(1 + \mu/(\kappa_1 E_)}),$$

where $E_0 = P_E/\delta_E$. The trivial steady stable is (unsurprisingly) stable when $R_0 < 1$, and the non-trivial steady state is stable when $R_0 > 1$. Biologically speaking, the trivial steady state corresponds to host clearance of the pathogen, whereas the non-trivial steady state corresponds to a persistence of the infection within the host for a prolonged period of time.

Wilson further noted that the host's ability to clear infection will increase if greater macrophage engulfment of chlamydial particles occurs prior to host cell infection (μ/κ_1), if the cytotoxic immune response clears infected cells prior to lysis of the infected cell (γ/κ_2), or if the number of infectious chlamydial particles released by an infected cell is reduced (P). We revisit these host responses as we develop our own model.

5.2.1 Application of the deterministic modelling approach

Modelling of infections based upon this basic framework considered extensions of the system in order to build knowledge about the dynamics of chlamydial function. Wilson et al. [11] applied this model (with an extension to consider the

number of reticulate bodies within an infected cell) to an investigation of the effectiveness of the Th1 immune response to chlamydia infection and concluded that the Th1 immune response has an exponential effect on the number of secondary infected cells, which emphasises the importance of cell-mediated immunity in clearing an infection. Wilson and McElwain [12] considered the impact of anitbody attachment to the pathogen in its extracellular form to investigate the importance of humoral immunity in clearing and infection. Vickers et al. [9] considered the CD4+ and antibody response to an infection, and then modified the rates of these responses under repeat infections to demonstrate how variation in immunobiological parameters in response to repeat infections had important clinical and epidemiological implications. More realistic behaviour was represented with a spatial dimension (represented with a set of partial differential equations) [5]. Host response scenarios can be modelled by changing parameter values of a particular equations, which in turn yields insights into the long-term behaviour of the system under different conditions. However, this modelling approach does not capture the randomness of the initial phase of an infection.

5.3 Branching process

Let N_t be the number of infected host cells at time $t \geq 0$, with initial conditions $N_t = N_0$. At time t, for the ith infected cell, with $i = 1, \ldots, N_t$, let Y_t^i be a t-measurable random variable that represents the number of infected progeny a particular cell generates. We assume that Y_t^i is independent and identically distributed according to some distribution f that is not a function of time, nor of the current state of the process N_t. We then model the number of infected cells from previous incidences as

$$N_t = \sum_{i=1}^{N_{t-1}} Y_t^i \,.$$

This defines the basic single-type branching process [2].

The progeny distribution f has mean and variance μ and σ^2, respectively. These two moments characterise the main properties of the process. The trajectories

of the process either die out or explode, such that

$$\Pr(\{\lim_{t\to\infty} N_t = \infty\} \cup \{\lim_{t\to\infty} N_t = 0\}) = 1\,.$$

The mean number of progeny μ acts as the bifurcation parameter of the process, in that $\mu \leq 1$ implies the almost sure extinction of the process. As a point of interest, the deterministic model $N_t = \mu N_{t-1}$ has the same bifurcation parameter. However, in the deterministic model the population persists to $N_t = N_0\,$. The stochastic model with $\mu = 1$ goes extinct almost surely.

In the supercritical case $\mu > 1\,$, the probability of the extinction of the process is given as the fixed point $q = f(q)\,$, where $f(z) := \mathbb{E}(z^{Y_t^i})\,$. That is, the probability generating function of the infectious progeny distribution gives an expression for the probability of extinction of an infection.

For the critical and sub-critical cases, define T to be the extinction time of the process. Then $N_T = \sum_{k=0}^{T} N_k$ follows a power series distribution when Y_t^i follows a power series distribution. That is, $\Pr(Y_t^i = k) = \alpha_k \lambda^k$. In particular, if the progeny random variable follows a Poisson distribution with parameter λ, then the total infectious load follows a Borel–Tanner distribution with parameters λ and N_0 [3].

5.4 Modelling infection outcomes

The branching process model tracks the randomness in the size of each burst, along with the chances that these bacteria go on to infect another cell. The primary entities in this setup are the number of infected host cells. Changes in the amount of extracellular infectious material are represented in the model to support a mechanistic interpretation of the results.

The model is generated with the birth of a new generation following a Poisson process with rate parameter $\lambda = 0.25$ where time is measured in hours. When a lysis or extrusion event occurs, a random draw is made from a Binomial distribution with parameters $n = 200$ (n is measured as number of extracellular bacteria), $p = 0.004\,$, (in units of new cells per extracellular bacteria) to represent

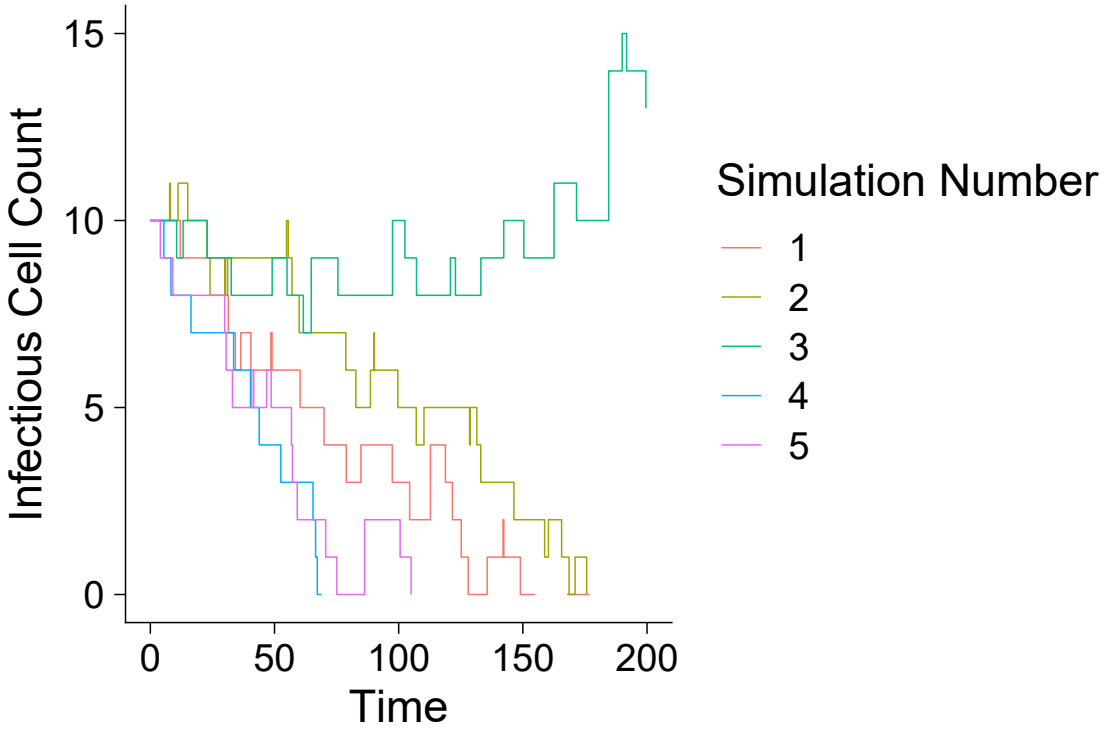

Figure 5.1: **Simulation of a branching process.** Event times are according to a Poisson process and with a binomially distributed offspring distribution.

the burst size and chance of each bacteria infecting a new cell, respectively. The random draw quantity is then added to the total number of infected cells, with one subtracted to represent the original cell's demise, so that there is a significant chance of no infectious progeny being produced and the population total decreasing in size. The value of parameters in these simulations were chosen for illustrative purposes, but are mostly consistent with other models in the literature [12, 5]. The parameters chosen for burst size and infection probability give an expected number of offspring per generation of -0.2 cells.

5.4.1 Given an initial infectious load, what are the chances the infection will develop?

In Figure 5.1 we observe the qualitative difference between realisations of a branching process under the same parameter conditions. Simulation four dies out almost immediately, whereas simulation three has a similar infected cell counts to the initial starting population. It is interesting to note that while the survival times

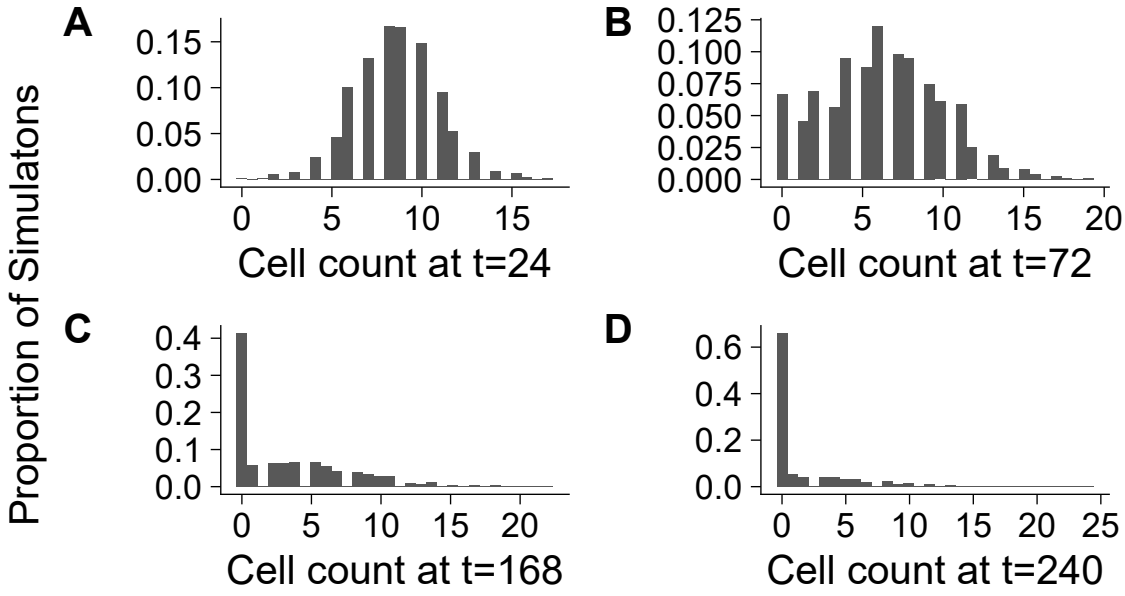

Figure 5.2: **A comparison of cell count distributions.** For stopping times of (A) 24, (B) 72, (C) 168 and (D) 240 hours.

in simulations one and two are approximately similar, the total cell count over time is significantly greater in two than in one. This represents a scenario in which the infectious burden is greater for one realisation compared to another, conditional on survival until a particular time.

5.4.2 Given an initial infectious load, what are the chances the infection will last until a given stopping time?

In Figure 5.2 we simulate the branching process under the same conditions as in Figure 5.1. We expect that the proportion of zeroes is a function of the stopping time since the zero is an absorbing state of the process. At time 72 less than a tenth of processes have a cell count of zero, whereas a majority have a cell count of zero at time 240. Alternatively, the closer the stopping time to the initial time, the more processes we expect to observe with a positive population count. Figure 5.2 shows the change in the distribution of cell counts for varying stopping times.

We count the number of cells at a stopping time $t = 240$ as this represents sufficient time for an infection to clear in most cases. The simulation is repeated 1000 times, to give an empirical density of the population count at a stopping

time. The histogram in Figure 5.2(D) shows the normalised density. This histogram shows that the distribution of cell counts is dominated by zeroes. In this simulation approximately two-thirds of the simulations had gone extinct by the stopping time.

The other implication of this example is that the distribution of positive cell counts have a long right-tail. Although we expect a significant proportion of processes to have a cell count of zero by a defined stopping time, there is still some significant probability of observing a large population count at the same stopping time. These unlikely but possible events may explain in part why some infections cause no observable harm, whilst others persist.

5.4.3 Given an initial infectious load, what are the chances the infection will be above a certain threshold at a given stopping time?

Whilst the previous question is concerned with the quantity $\Pr(N_T \mid t = T)$, this question asks $\Pr(N_T > N^* \mid t = T)$ for any arbitrary threshold $N^* \geq 0$. We would like to know not only if an infection will be present at a certain point in time, but if it will be present in large enough quantities to cause inflammatory damage. This is crudely addressed with each of the subfigures in Figure 5.2. For a particular stopping time, the proportion of cell counts above a particular threshold corresponds to a region under the empirical distribution. The figures shows that the smaller stopping times are seen to correspond to a greater region of probability density compared to larger stopping times.

5.4.4 Given an initial infectious load and subcritical/critical reproduction, when will extinction occur?

We are interested in the time to extinction of a branching process with critical or subcritical reproduction, as this is a factor for total infectious load. A longer time to extinction will imply a greater infectious load. It is a straightforward consequence of the results above that most processes will go extinct in a short amount of time. We observe the existence of a long right-tail in the distribution of

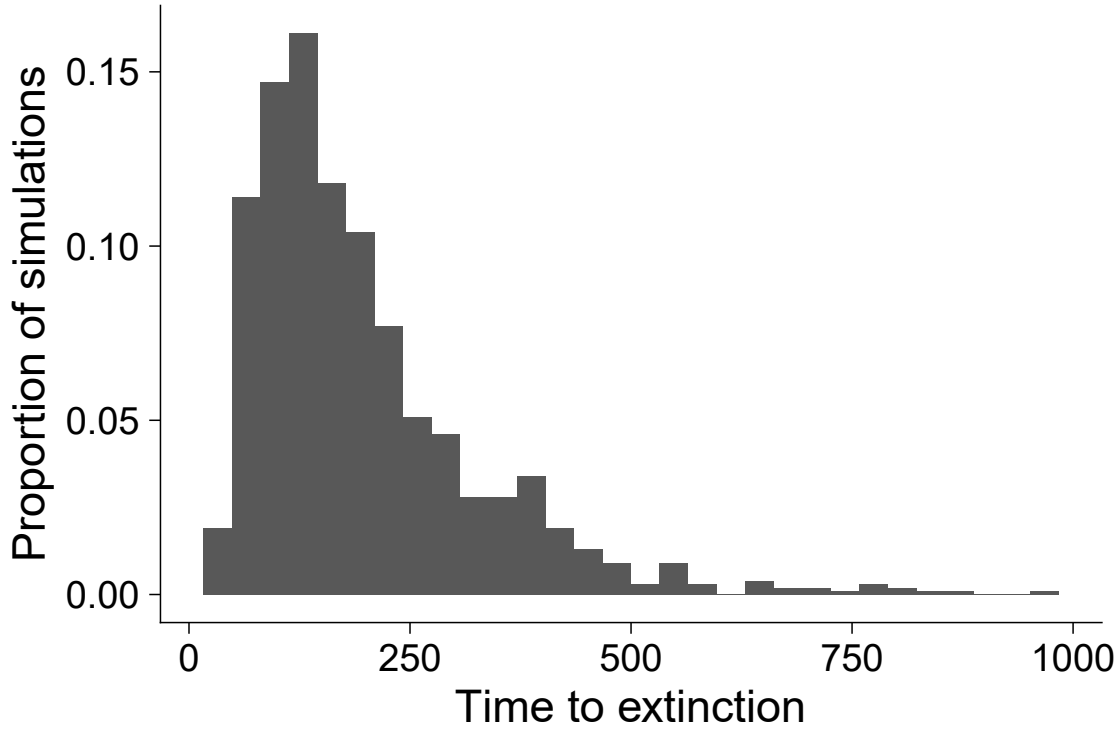

Figure 5.3: **Histogram of time to extinction.**

extinction times in Figure 5.3.

5.4.5 Given an initial infectious load and subcritical/critical reproduction, what will the total infectious load be?

The total infectious load is defined as the total cell population count multiplied by the length of each cell's lifetime. It is the metric that gives the best sense of how severe a particular infection may be, as inflammation due to the immune response is a function of total infectious load. Figure 5.4(A) shows the total infectious load, which is a similar density type to that of the extinction time, with a majority of processes causing a small amount of total load and a long right-tail.

Figure 5.4(B) also shows the relationship between the time to extinction and total infectious load. The two are linearly related, since total infectious load is defined as a the product of extinction times and population count. It is interesting to note that there is a large degree of variability around this mean relationship. There exist some cases where the stopping time of the process is short, but the

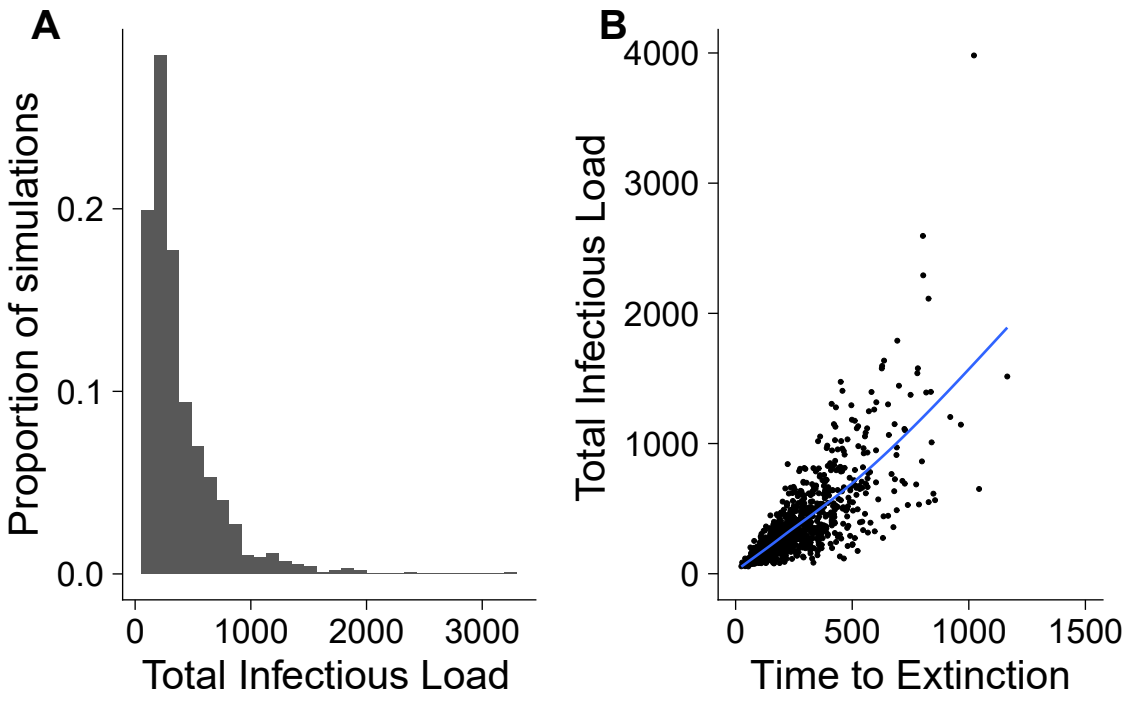

Figure 5.4: **Summarising total infectious load and time to extinction.** (A) Histogram of total infectious load; and (B) The relationship between the time to extinction and total infectious load, where the plotted line is least squares best fit.

total infectious load is quite high. Conversely, there are some processes where the extinction time is long but the total infectious load is quite low.

In all of the above examples, it appears that a majority of processes will result in a small impact for the given measure of interest (stopping cell count, extinction time, total infectious load). However, for a number of processes this will not be the case, and the impact (for the given measure of interest) will be large. When considering the chances of a particular infection to progress to further disease, we recognise that most infections do not progress to disease at all, but those that do may have a large impact.

5.5 Modelling the immune system to investigate sensitivity to parameters in the model

We map different responses of the immune system onto parameters of the model. For example, a larger adaptive immune system response may reduce the

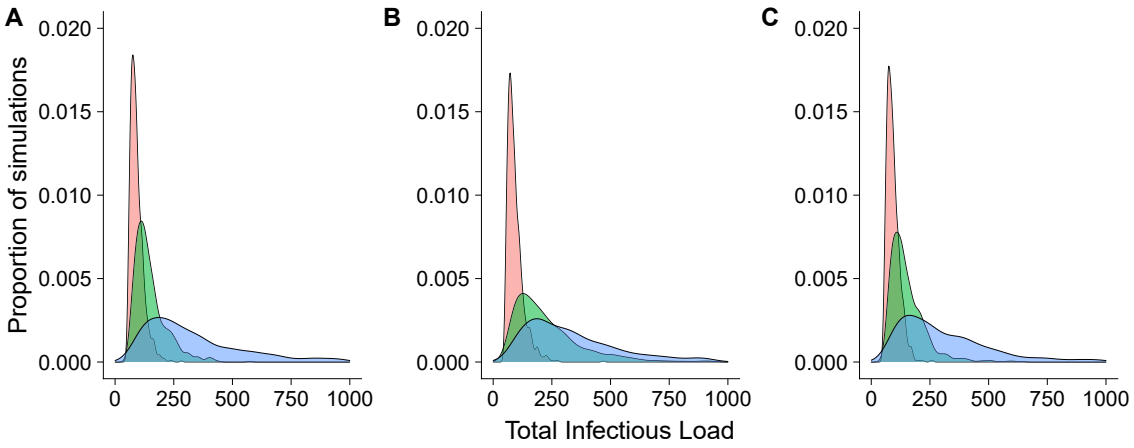

Figure 5.5: **Investigating immune system influences on the model.** Density of total infectious load for three different immune system responses, where the red densities represents a high immune response, green a medium response and blue representing no immune response. (A) Macrophage Engulfment; (B) Clearance prior to lysis; (C) Reduction in burst size.

tail of the progeny distribution. This allows us to use our model to determine how particular choices of parameters plays a role in modifying particular outcomes, such as the stopping time of a sub-critical infection or the total infectious load, and how the immune system in the real-world process might influence these outcomes. We describe three specific immune system responses and the outcomes of the model they produce. Immune responses, represented by different choices of parameter values, are compared for the total infectious load of a particular realisation of an infection, as this best represents the associated burden of the infection and the progression to pathology.

5.5.1 Macrophage engulfment

Macrophages are components of the innate immune system that migrate within tissue and detect the presence of pathogens. Chlamydia are cleared by macrophage engulfment prior to host cell infection [10]. We model this immune system response by varying the success probability of an extra-cellular bacteria infecting a new host cell. Figure 5.5(A) plots the total infectious load (as defined above) for three scenarios, where $p = 0.04$ representing no immune response of this

type and $p = 0.03$, $p = 0.02$ representing medium and high responses of this type, respectively.

A high response of this type (distribution in red in Figure 5.5(A)) is effective in reducing the total infectious load or burden close to zero for most infections. Most infections are also rapidly cleared for a medium response (distribution in green), however there is a significant proportion of infections with a higher infectious load.

5.5.2 Cell-mediated immune response

The T helper cells, particularly TH1 cells (also known as CD4+ cells) are components of the adaptive immune system that are capable of detecting damaged cells of the host, which increases the host's ability to clear an infection by removing an infected cell prior to lysis or extrusion occurring [11]. This is equivalent to the burst size of the infected cell being equal to zero. We introduce an extra parameter p_0 to our model that represents the probability a cell will be cleared prior to lysis or extrusion. To model medium and high immune responses, we simulate the total infectious load for $p_0 = 0.1, 0.5$.

Similarly to the macrophage response, a high response of the cell-mediated immune type is effective at clearing the infection rapidly. However, unlike the macrophage response, a medium response may still allow a number of infections to have a larger infectious load.

5.5.3 Burst size

The burst size may also be impacted by a varying immune response [12]. We model this by first assuming that the burst size of an infected cell is drawn from a Poisson distribution with parameter $\lambda = 200$, as opposed to a fixed burst size as above. To represent a medium and high immune response of this type, we then modify the mean of the burst distribution to $\lambda = 150, 100$, respectively.

Figure 5.5 shows a comparison between no response, medium and high

responses of each of the three types described above. This exercise constitutes a sensitivity analysis of the model; however, it does give some sense of the impact of the response by each component of the immune system. It should be noted that each response is considered independently of the others and assumed constant over time, which is a biological oversimplification.

The comparison above demonstrates that for this model, an effective immune response that neutralises free extra-cellular particles has the greatest impact on the distribution of total infectious load, as the behaviour of the model is most sensitive to changes in the success probability parameter. This can be observed by comparing the high and medium response in Figure 5.5(C) to those displayed in (A) in (B).

Overall, this section demonstrates the influence of differing values of parameters of the model. It confirms that the overall distribution of outcomes, and in particular the skew of the distribution, is produced over a range of choices for the parameters. The overall reduction in total burden may suggest that host immunological factors may be significant in determining the progression to pathology for an individual, if the correspondence between the immune system and parameters of the model can be demonstrated empirically.

5.6 Discussion

In this chapter, we have set up a stochastic model that represents the progression of infection using branching processes. The unique biphasic developmental cycle of Chlamydia was represented in the progeny distribution for each generation of infected cells.

This work can be thought of as extensions to previous deterministic mathematical models of Chlamydia. The inclusion of stochastic elements demonstrates the influence of randomness on the survival or extinction of a particular infection. Randomness in the model also generates a long-tailed distribution of total infectious load. Although this long-tail does not directly predict pathology, it may be a good correlate, and therefore explains why some infection progress to pathology whilst

others do not. A key piece on any future modelling of infection processes will be to investigate which factors may increase the skew of the distribution of infectious load, or predict higher infectious load compared to a population distribution.

As would be expected, factors (such as immune responses) that reduce the size of the progeny distribution of infected cells were more influential in preventing further infectious load. It was of interest that certain components were more effective than others are reducing total infectious load, particularly at lower levels of the type of the immune response.

Branching processes provide useful tools to evaluate outcomes of the model. The probability generating function of the progeny distribution can be solved to find the probability of extinction, for the case that the mean of the distribution is greater than 1. This concept could be extended to more complex distributions by finding the Kolmogorov forwards or backwards differential equation for the probability generating function, and finding an analytical or numerical solution. Overall this class of models can be easily extended to represent particular biological phenomena, as will be shown in the next chapter.

5.7 Concluding remarks

In this chapter we consider a class of stochastic processes for modelling chlamydial infections, and show that realisations of a branching process that has a mechanistic definition reproduce biological phenomena. We demonstrate how this class of models permit large variability in outcomes, and discuss how this may be influential in the progression to pathology. The distribution of outcomes has a long right-tail, which suggests that pathology is not an average outcome based on our investigation of the dynamics of the underlying process. The model is a significant first step in work to best elucidate the mechanism by which an infection will result in severe reproductive sequelae, and determining the various host and pathogen factors that moderate pathology.

References

[1] Yasser M. AbdelRahman and Robert J. Belland. "The chlamydial developmental cycle". In: *FEMS Microbiology Reviews* 29.5 (Nov. 2005), pp. 949–959. ISSN: 0168-6445. DOI: 10.1016/j.femsre.2005.03.002.

[2] T. E. Harris. *Branching Processes*. Vol. 19. 4. The Institute of Mathematical Statistics, Dec. 1948, pp. 474–494. DOI: 10.1214/aoms/1177730146.

[3] Christine Jacob. "Branching processes: Their role in epidemiology". In: *International journal of environmental research and public health* 7.3 (Mar. 2010), pp. 1186–1204. DOI: 10.3390/ijerph7031204.

[4] N Low et al. "Incidence of severe reproductive tract complications associated with diagnosed genital chlamydial infection: the Uppsala Women's Cohort Study". In: *Sexually Transmitted Infections* 82.3 (2006), pp. 212–218. ISSN: 1368-4973. DOI: 10.1136/sti.2005.017186.

[5] Dann Mallet et al. "A Mathematical Model of Chlamydial Infection Incorporating Movement of Chlamydial Particles". In: *Bulletin of mathematical biology* 75 (Oct. 2013). DOI: 10.1007/s11538-013-9891-9.

[6] Heather K. Maxion et al. "The Infecting Dose of Chlamydia muridarum Modulates the Innate Immune Response and Ascending Infection". In: *Infection and Immunity* 72.11 (2004), pp. 6330–6340. ISSN: 0019-9567. DOI: 10.1128/IAI.72.11.6330-6340.2004.

[7] S. Menon et al. "Human and Pathogen Factors Associated with Chlamydia trachomatis-Related Infertility in Women". In: *Clinical Microbiology Reviews* 28.4 (2015), pp. 969–985. ISSN: 0893-8512. DOI: 10.1128/CMR.00035-15.

[8] Martin Nowak and Robert May. *Virus Dynamics: Mathematical Principles of Immunology And Virology*. Oxford University Press, Jan. 2001.

[9] David M. Vickers, Qian Zhang, and Nathaniel D. Osgood. "Immunobiological Outcomes of Repeated Chlamydial Infection from Two Models of Within-Host Population Dynamics". In: *PLOS ONE* 4.9 (2009), e6886. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006886. URL: https://dx.doi.org/10.1371/journal.pone.0006886.

[10] D.P. Wilson and D.L.S. McElwain. "A model of neutralization of Chlamydia trachomatis based on antibody and host cell aggregation on the elementary body surface". In: *Journal of Theoretical Biology* 226.3 (2004), pp. 321–330. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2003.09.010.

[11] D.P. Wilson, P. Timms, and D.L.S. McElwain. "A mathematical model for the investigation of the Th1 immune response to Chlamydia trachomatis". In: *Mathematical Biosciences* 182.1 (2003), pp. 27–44. ISSN: 0025-5564. DOI: https://doi.org/10.1016/S0025-5564(02)00180-3.

[12] David Wilson. "Mathematical Modelling of Chlamydia". In: *The ANZIAM Journal* 45 (Apr. 2004). DOI: 10.21914/anziamj.v45i0.883.

6

Ascension of Chlamydia is moderated by uterine peristalsis and the neutrophil response to infection

6.1 Introduction

There is considerable evidence to indicate that *Chlamydia trachomatis* in women is one of the etiological agents responsible for Pelvic Inflammatory Disease (PID), and that women who experience PID are more likely to experience tubal factor infertility and ectopic pregnancy [19, 14, 25, 7]. Laparoscopic and histological or molecular evidence has indicated that infection can be present in the fallopian tubes, even when non-detectable at the cervix, the primary site of infection [23]. It is suggested that the development of tubal pathology is a result of a chlamydial infection ascending into the upper reproductive tract combined with a damaging pro-inflammatory response to the pathogen [18].

While the proportion of infections that ascend are not known, there are estimates on the frequency of damaging pathology. In the POPI trial, 9.5% women with a chlamydial infection developed PID [19]. A longitudinal study of 1844 women with laparoscopically confirmed PID (all aetiologies) found that approximately 30% experience tubal factor infertility [25]. Ascension of *Chlamydia trachomatis* into the upper genital tract can occur without necessarily developing pathology. The

relatively low rate of pathology development in women with infections [9] suggests that a number of host and pathogen factors moderate ascension, such as menstrual cycle, stage of infection, immune status (and other influential factors not considered in this study, such as genotypic factors (host and pathogen), cellular exit strategies of the pathogen (lysis or extrusions) and host microbiome composition). The infectious load (the amount of infection when the infection is first received) may also determine the likelihood of ascension. Maxion et al. described how the infectious load of *Chlamydia muridarum* in mice was more likely to result in viable Chlamydia in the oviduct when the animals received lower infecting doses [17]. Higher infecting loads appeared to result in an immune profile in favour of cells associated with the innate immune response. This study found the response of immune cells (innate and adaptive) was independent of location (lower or upper genital tract) for the relationship with infecting load. The findings from this study suggested that innate immune cells contribute to the control of ascending infection, such that the immune response for high infectious loads of Chlamydia was sufficient to prevent ascension into the upper genital tract, but not so for lower loads.

A comprehensive review identified an association between infectious load, specimen type, and anatomical site of infection [24]. However, the evidence was unclear for the role of infectious load in ascending infection. An experiment on the effects of infectious load in female mice found that higher inoculating doses ascended to the upper genital tract more rapidly [4]. However, in the experimental protocol, mice were given 2.5g of medroxyprogesterone acetate 7 days before infection, and so the interplay between initial load and timing of the menstrual cycle was not evaluated. Only one study [1] looked at the relationship between load at the cervix and PID diagnosis, which found an association between higher load and PID. However, the study design was cross sectional, and so a causal interpretation cannot be made.

All chlamydiae share a unique biphasic developmental cycle, where infectious extracellular elementary bodies (EBs) enter cells and differentiate into replicative reticulate bodies (RBs). After several rounds of replication, RBs differentiate back into EBs and are released from the host cell, in order to infected neighbouring cells [1]. Each host (humans, guinea pig, mice etc.) generally has a distinct species

of Chlamydia, and there are 12 species currently identified but the overall biphasic developmental cycle process is consistent throughout the genus. Chlamydia does not appear to encode a specific mechanism for movement on its genome. A one-dimensional spatiotemporal model described the population dynamics of chlamydial particles, infected and uninfected epithelial cells, and with movement that was determined by a diffusion coefficient in an idealised representation of the cervix, with random diffusion at about 10-4 $m^2 s^{-1}$ (with an uncertainty of an order of magnitude) in the elementary body form (EB) [16, 15]. The model demonstrated that the peak infection load and time to clearance was a function of the movement of *Chlamydia trachomatis*, and that increased motility may lead to stronger host immune response. However, the scale of the diffusion term, in comparison to the length of the cervix implied that diffusion of the elementary bodies is an implausible process for ascension.

A model for movement or ascension has been proposed by which polymorphonuclear leukocytes (PMNs) infiltrated the conjunctival epithelium following a guinea pig infection with *Chlamydia caviae* [21]. The response was in the early stages of the infection, which concords with the previously described critical role of neutrophils in controlling the early stage of an infection [2]. The PMN response was observed to result in detached superficial epithelia from the mucosa, including infected cells and elementary bodies that remained healthy and intact during the process. It was suggested by Rank et al. that this response was a defensive mechanism from the perspective of the host organism in that it allows for infected cells that have been ejected from the epithelium to be cleared from the infection site. Further, it was proposed that from the perspective of the Chlamydia, this process allowed for movement away from infected sites where there is a lack of susceptible uninfected cells, transporting to new infection sites via fluid dynamics or tissue movement. Movement and secretion of materials via fluid dynamics in the genital tract is governed by uterine peristalsis. The interplay of peristaltic contractions and ascension of sexually transmitted infections has not been examined and represents a crucial gap in current understanding.

It has been established that there is bi-directionality of the uterine peri-

stalsis dependent on the stage of the menstrual cycle. In work conducted by Kunz and Leyendecker (2001), vaginal sonography of uterine peristalsis in women has been used to characterise the dynamics of rapid sperm transport [12]. They found an increase in the frequency and intensity of subendometrial and myometrial peristaltic waves during the follicular phase of the menstrual cycle. The number of waves in the fundo-cervical direction decreased and waves in the cervico-fundal direction increased at this stage of the cycle, suggesting that sperm transport is facilitated by these uterine conditions, and that the extent of migration is increased with the progression of the follicular phase.

It is therefore plausible that the immune response which detaches infected cells from the epithelium and into the cervical canal can result in the transport of infected cells and elementary bodies via peristalsis into higher regions of the genital tract. Since the direction and intensity of peristaltic movement is dependent on the menstrual cycle, the stage of the menstrual cycle at infection could be one governing factor in determining which women experience infections that ascend or descend the genital tract, although it may be further impacted by the infectious burden (as explored in this model)

The phenomena described above could be combined to provide a process for infective chlamydial bodies to enter into the genital tract and be transported to the upper reproductive tract via peristalsis. From there, the infection cycle described above could start, with pro-inflammatory immune responses following the infection sites to the upper genital tract. In this work, we represent this process with stochastic model of the dynamics of a chlamydial infection, the associated host response, and the movement of infectious material. This allows us to better appreciate the complex, dynamic, and nonlinear interactions between the neutrophil response, the infectious load, and the menstrual cycle that moderate chlamydial ascension. Elucidating the processes involved in chlamydial ascension will provide a better understanding of the pathogenesis of *Chlamydia trachomatis* associated PID and infertility. In turn, the moderating factors could enable more specific testing for at-risk women, or development of other interventions when infections are diagnosed.

6.2 Methods

6.2.1 Developing a stochastic model to determine the relative influence of factors that moderate ascension

We developed a stochastic model of Chlamydia trachomatis and infected cells in two compartments; the cervical epithelium and cervical canal. The structure of the model is informed by previous mathematical modelling studies of within-host chlamydial dynamics [27, 16, 15], and by synthesis of the literature on the chlamydial developmental cycle and host cell interactions. The model developed for this study extends previous models in order to incorporate elements specific to ascension (such as peristalsis, neutrophil ejections, epithelial and cervical elements) and stochastic elements to incorporate uncertainty. Figure 6.1 displays the host cell interactions that we represent in our model in order to evaluate the moderating impact of each on ascension.

We infer parameters for our model from plausible prior distributions (Table **??**). In turn, prior distributions are parameterised where possible, using values from previous modelling studies. Transitions are stochastic meaning realisations of the model permit variable outcomes. This allows us to infer behaviour about an infection over a broad range of plausible parameter values, including phenomena not explicitly represented in the model.

Finally, the simulations are classified by a binary condition that reflects if the population of infectious material in the cervical canal compartment is non-zero at the endpoint of the simulation. If there is infectious material present (i.e. non-zero in the cervical canal), then ascension is said to occur in that iteration. This outcome is then used to determine relationships between the parameters of the process and the proportion of simulations where ascension occurred (reflecting peristalsis towards the fundus and live infectious material in the cervical canal). Relationships between the parameter values and the outcome are graphically summarised using natural splines. Each simulation is begun with a fixed amount of time remaining before peristaltic contractions move fluid in the cervical canal upwards, causing any infectious material

to ascend. We stop each simulation at the moment when peristalsis causes material to begin ascending, with the assumption that any material in the cervical canal at this point will ascend.

6.2.2 Defining the process

There are two entities of interest in this modelling exercise, extracellular chlamydial elementary bodies (EBs) and infected cells containing chlamydial reticulate bodies (RBs) and EBs [1]. We denote $X_{\{C_1,t\}}$, $X_{\{C_2,t\}}$, $Y_{\{C_1,t\}}$, $Y_{\{C_2,t\}}$ as random variables denoting the number of extracellular chlamydia (represented as X) and infected cells (represented as Y) in the epithelium and cervical canal respectively (represented by the subscripts C_1,C_2).The model assumes that infected cells produce extracellular bacteria in an instantaneous burst, as opposed to a continuous process, which is important to note as this choice may impact the behaviour of generated quantities of the model [20].

Chlamydia trachomatis has two distinct exit strategies from a cell, lysis and extrusions, which are mediated by distinct mechanisms [10]. Extruded chlamydial inclusions may exit the cell slower than what occurs via lysis in the rate of exit, and the amount of *Chlamydia trachomatis* released from the cell may differ (albeit this difference is likely subtle compared to the other parameters we explore here). Further, these factors governing chlamydial cell exit may vary by serovar type and various host specific factors. Therefore, we utilise broad parameter ranges over the function of extracellular bodies in order to represent the variation attributable to these factors without explicitly representing them. To this end, there is only one type of extracellular *Chlamydia trachomatis* represented in the model, although we recognise that there are many other potentially influential factors in the process.

We utilise the framework of previous models of chlamydial infections by Wilson to develop our model [29]. The realised outcomes of a particular sample path in a model of *Chlamydia trachomatis* is proposed to be strongly influenced by the random elements [3], hence our model incorporates stochasticity. Representing the process as a stochastic model allows us to represent the uncertainty attributable

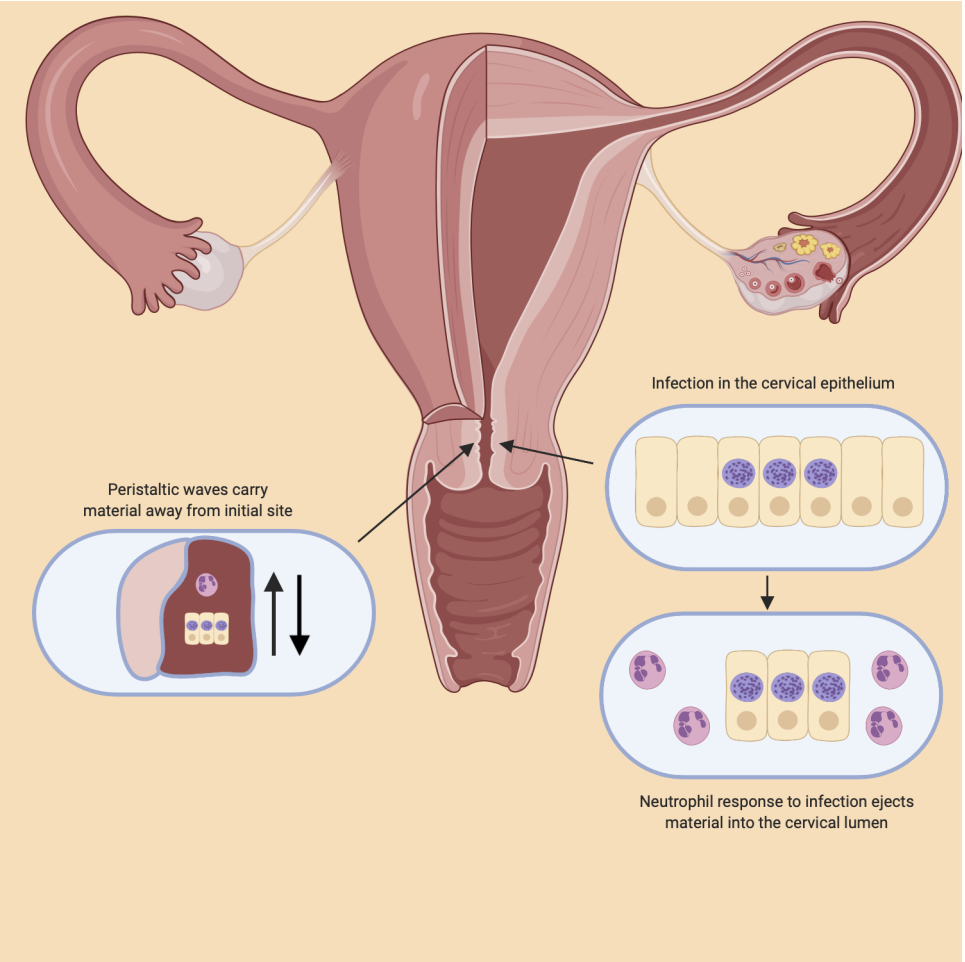

Figure 6.1: **The proposed model of chlamydial ascension.** The figure depicts the female reproductive tract and elements that were evaluated in their moderating impact of ascension. The factors evaluated were the chlamydial cell cycle, the host and pathogen interactions including the immune system response and the ejection of infectious material from the epithelium, and the effect of peristalsis in moving material either away or towards the upper reproductive tract depending on the stage of the menstrual cycle.

to other factors that are believed to be important in moderating infections, such a host and pathogen genotype and the cervico-vaginal microbiome.

For the ascension model, we extend the dynamics of the process to include ejection from the epithelial layer of the cervix into the cervical canal. We denote with the subscript C_1 for infected cells and EBs in the epithelium, and C_2 for infectious units in the cervical canal. The scheme of the process can be represented as follows (following the convention of Pearson et al. [20].

$$
\begin{aligned}
X_{C_1} &\to \emptyset \text{ rate } \gamma_1 \ (EBs \ per \ minute) \\
Y_{C_1} &\to \emptyset \text{ rate } \gamma_2 \ (infected \ cells \ per \ minute) \\
IX_{C_1} &\to Y_{C_1} \text{ rate } \kappa_1 \ (EBs \ per \ minute) \\
Y_{C_1} &\to BX_{C_1} \text{ rate } \kappa_2 \ (infected \ cells \ per \ minute) \\
M_1X_{C_1} &\to M_1X_{C_2} \text{ rate } \alpha \ (EBs \ per \ minute) \\
M_2Y_{C_1} &\to M_2Y_{C_2} \text{ rate } \alpha \ (infected \ cells \ per \ minute) \\
X_{C_2} &\to \emptyset \text{ rate } \gamma_3 \ (EBs \ per \ minute) \\
Y_{C_2} &\to \emptyset \text{ rate } \gamma_3 \ (infected \ cells \ per \ minute) \\
Y_{C_2} &\to Y_{C_1} \text{ rate } \kappa_3 \ (infected \ cells \ per \ minute)
\end{aligned}
$$

The vector $(X_{\{C_1,t\}},\ X_{\{C_2,t\}},\ Y_{\{C_1,t\}},\ Y_{\{C_2,t\}})$ represents the current state vector of the process. The empty set $\emptyset$ is used to denote an infectious body (EB or infectious cell) that has been removed by the immune system or peristaltic movement. For this model, we have possible transitions that correspond to the removal of infectious material: $(-1, 0, 0, 0)$ representing the death of an extracellular EB, $(0, -1, 0, 0)$ representing the death of an epithelial infected cell, and $(0, 0, -1, -1)$ representing the removal of EBs and infected cells in the cervical canal.

$I(t)$, $B(t)$ are random variables that determine the number of bacteria required to form an infectious unit, and the number of bacteria formed by a bursting event. $M_1(t)$, $M_2(t)$ are random variables describing the ejection from the epithelium

Parameter	Meaning	Distribution	Value	Supporting Literature
$\gamma_{1,innate}, \gamma_{1,adaptive}$	Death rate of elementary bodies (EBs) (per minute) due to the innate and adaptive immune response	Truncated Gamma	1, 100, τ_1	[16, 29, 28, 22]
$\gamma_{2,innate}, \gamma_{2,adaptive}$	Death rate of infected cells (per minute) due to immune response	Truncated Gamma	1, 100, τ_2	[16, 29, 28, 22]
τ_1	Threshold of extracellular chlamydia at which immune system increases	Negative Binomial	5000, σ_1	Assumed
τ_2	Threshold of infected cells at which immune system increases	Negative Binomial	150, σ_2	Assumed
σ_1	Variance of threshold parameter	Beta	1, 1	Assumed (hyperparameter)
σ_2	Variance of threshold parameter	Beta	1, 1	Assumed (hyperparameter)
κ_1	Rate of inclusions of EBs (per minute) into new infected cells	Gamma	1, 100	[16, 29, 28, 22]
κ_2	Rate of burst of infected cells (per minute) to new EBs	Gamma	1, 100	[16, 29, 28, 22]
α	Rate of migration of infectious material due to the neutrophil response	Gamma	1, 100	Informed by [21]
I	Number of new inclusions of infected cells	Negative Binomial	10, σ_4	[16, 29, 28, 22]
B	Number of new EBs in each burst event	Negative Binomial	350, σ_3	Mean and dispersion chosen to match the 200-500 nominal range as per [16, 29, 28, 22]
M_1	Number of chlamydia per migration event	Negative Binomial	100, σ_5	Informed by citerank2008
M_2	Number of infected cells per migration event	Negative Binomial	10, σ_6	Informed by [21]
σ_3	Variance of inclusion size	Beta	2, 2	Mean and dispersion chosen to match the 200-500 nominal range as per [16, 29, 28, 22]
σ_4	Variance of the burst size	Beta	2, 2	Assumed (hyperparameter)
σ_5	Variance of the migration for EBs	Beta	1, 1	Assumed (hyperparameter)
σ_6	Variance of the migration for infected cells	Beta	1, 1	Assumed (hyperparameter)
$\gamma_{3,fixed}$	Rate of removal (per minute) of infectious material due to the intensity of peristaltic contractions	Gamma	1, 1	Assumed (unique to this study)

Table 6.1: **Parameters incorporated in the model and values assigned for this study.** Details on the parameterisation used for each of the distributions in the model can be found at the end of the methods section.

to the cervical canal of EBs and infected cells respectively.

The trading between infectious material in the epithelium is characterised by $(-I(t), 1, 0, 0)$, where $I(t)$ EBs from an inclusion to produce one infected cell. Conversely, infected cells can burst to produce $B(t)$ new EBs, represented by $(B(t), -1, 0, 0)$. Ejection is represented by $(-M_1(t),\ -M_2(t),\ M_1(t),\ M_2(t))$. Infected cells in the cervical canal can burst and produce EBs, represented by $(0, 0, B(t), -1)$.This final transition does not explicitly represent extrusions, however the parameters that govern the function of extracellular *Chlamydia trachomatis* are allowed to vary over ranges that could include this form of *Chlamydia trachomatis*.

6.2.3 Simulating the model

We simulate the model in continuous time, where time is measured in minutes. We use a grid of initial conditions for $X_{\{C_1,t=0\}} = x_0$ (from 1 to 10,001 in steps of 100, with 10 simulations at each value) and stage of the menstrual cycle (from 0 to 43,200 minutes in steps of 1800, with 10 simulations at each value) for a total of 252, 500 simulations, and assume that $X_{\{C_2,t=0\}} = Y_{\{C_1,t=0\}} = Y_{\{C_2,t=0\}} = 0$. Events occur according to a Poisson process, with rate parameter λ_t being the sum of all rate parameters as described in the process scheme above. Figure 6.2 describes the transitions between states.

The parameters γ_1 and γ_2 that govern the immune response are made up of

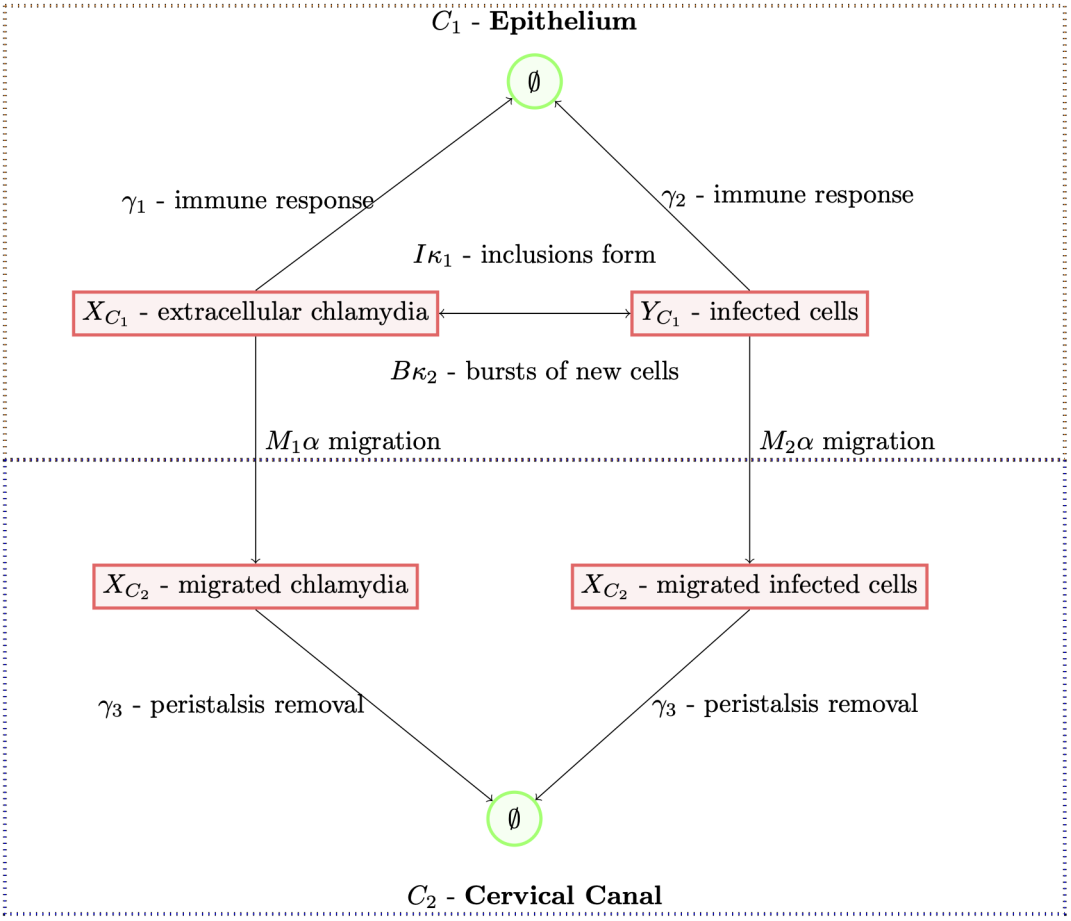

Figure 6.2: **Flow diagram of the model.** The figure outlines each component of the model that are incorporated in the analysis. Extracellular chlamydia can form inclusions to created new infected cells, and infected cells can burst to create new extracellular chlamydia. Both extracellular chlamydia and infected cells can be removed by the immune system, and be migrated to the cervical canal via the neutrophil response. In the cervical canal, infected cells can burst to create new chlamydia, but not vice versa. Extracellular Chlamydia and infected cells in the cervical canal be removed via peristalsis. The parameters of each component from the model are drawn from distributions that represent host and pathogen specific factors. The impact of each factor is represented as the proportion of simulations where the infection ascended for each parameter in the model in the following figures, out of a total of 252,500 simulations that were run.

two components, one representing the innate immune system and one representing the adaptive immune system. The adaptive immune response is assumed to operate once a certain threshold level of each infectious unit is achieved and is zero otherwise. These two thresholds (one for EBs, one for infectious cells), are drawn from a negative binomial distribution, with mean 5000 for EBs and 150 for infected cells. The shape parameter for each threshold is drawn from a $Beta\,(1,\ 1)$ distribution. It is assumed that the immune system response for extracellular bacteria is independent to the response for infected cells. The immune portion of the model is as follows

$$\gamma_1 = \gamma_1, innate + \gamma_1, adaptive$$

$$\gamma_2 = \gamma_2, innate + \gamma_2, adaptive$$

$$\gamma_1, innate,\ \gamma_2, innate \sim Gamma\,(1, 100)$$

$$\gamma_1,\ adaptive\ \sim\ 0\ if\ X_{C_1} \leq\ \tau_1,\ Gamma\,(1, 00)\ if\ X_{C_1} > \tau_1$$

$$\gamma_2,\ adaptive\ \sim\ 0\ if\ Y_{C_1} \leq \tau_2,\ Gamma\,(1, 00)\ if\ Y_{C_1} > \tau_2$$

$$\tau_1 \sim NegBinom\,(5000, \sigma_1)$$

$$\tau_2 \sim NegBinom\,(150, \sigma_2)$$

$$\sigma_1, \sigma_2 \sim Beta\,(1, 1)$$

A consideration for the model was the unique nature of chlamydial growth inside an intracellular inclusion vacuole where the organism replicates in multitudes, and the release of the infectious particles in bursts (by extrusion or lysis). The rate of inclusion developments and bursts are denoted by κ_1 and κ_2. The burst and inclusion sizes are drawn from a negative binomial with means 350 and 10 respectively, where the shape parameter comes from a $Beta\,(2, 2)$ distribution. The chlamydial inclusion and burst portion of the model can be represented as follows

$$\kappa_1, \kappa_2 \sim Gamma\,(1, 100)$$
$$B \sim NegBinom\,(100, \sigma_3)$$
$$I \sim NegBinom\,(10, \sigma_4)$$
$$\sigma_3, \sigma_4 \sim Beta\,(2, 2)$$

The rate of migration of infectious material is denoted as α, which assumes that the rate of migration is the same for both EBs and infected cells (i.e. representing cells displaced by the neutrophil response as outlined in the introduction). The number of EBs and cells being migrated due to the neutrophil response is also drawn from a negative binomial, with mean 100 and 10 respectively. The shape parameter for this random variable is drawn from a $Beta\,(1,\ 1)$ distribution. All shape parameters in the model act as hyperparameters for the uncertainties on the model parameters. They are rarely observed in the literature, and so they are drawn from weakly informative distributions about the exact parameter value. This addition of uncertainty makes the model results less sensitive to specific choices of parameter values.

$$\alpha \sim Gamma\,(1, 100)$$
$$M_1 \sim NegBinom\,(100, \sigma_5)$$
$$M_2 \sim NegBinom\,(10, \sigma_6)$$
$$\sigma_5, \sigma_6 \ \sim \ Beta(1,\ 1)$$

The parameter $\gamma_3\,(t)$ represents the removal rate of material in the cervical canal at time t. It is modelled as a fixed component and a function that represents variation due to the effect of the menstrual cycle on peristalsis. This assumes that changes in peristalsis occur continuously across the menstrual cycle, and that the menstrual cycle lasts for 30 days (30 days was arbitrarily assigned, although the

model parameters allow for a broad range of variability given the 21-35 day range in the population). The period and phase shift of the sine function are based on the number of minutes in 30 days, and the range is restricted to the interval $[0, 1]$. Peristaltic movement is represented as follows

$$\gamma_3(t) \;=\; \gamma_{3,fixed} + \; 0.5 \; sin\left(\frac{\pi\,(t - 10800)}{21600}\right) \; + 0.5$$

$$\gamma_{3,fixed} \sim Gamma\,(1, 1)$$

The inclusion of probability distributions over the parameters governing the model can be thought of as including uncertainty of host and pathogen specific factors that are not otherwise explicitly modelled. Ascension is defined as having occurred within a simulation, if the population of extracellular *Chlamydia trachomatis* and infected cells, X_{C_2},Y_{C_2} is greater than zero when the simulation time finishes. We then summarise the relationship between parameter values of the process, and the proportion of simulations where ascension occurred. Relationships are summarised graphically in the results section using cubic splines.

6.2.4 Parametrisation of key distributions

The gamma distribution used to draw key rates for the model is parameterised in terms of a shape parameter α and a rate parameter β, such that the expected value is α/β, and the probability density function can be represented as

$$f\,(x) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma\,(\alpha)}$$

The negative binomial distribution, used for some key parameters of the model, is parameterised in terms of the number of trials before the n^{th} success, k and the probability of success on each trial, p. The probability mass function is

$$f\,(x) = \binom{x \;+\; k \;-\; 1}{k \;-\; 1}\,(1-p)^k\,p^x$$

The truncated gamma distribution referred to is similar to the gamma distribution, with the inclusion of a threshold value t, such that the support of the distribution is $(t, \infty]$

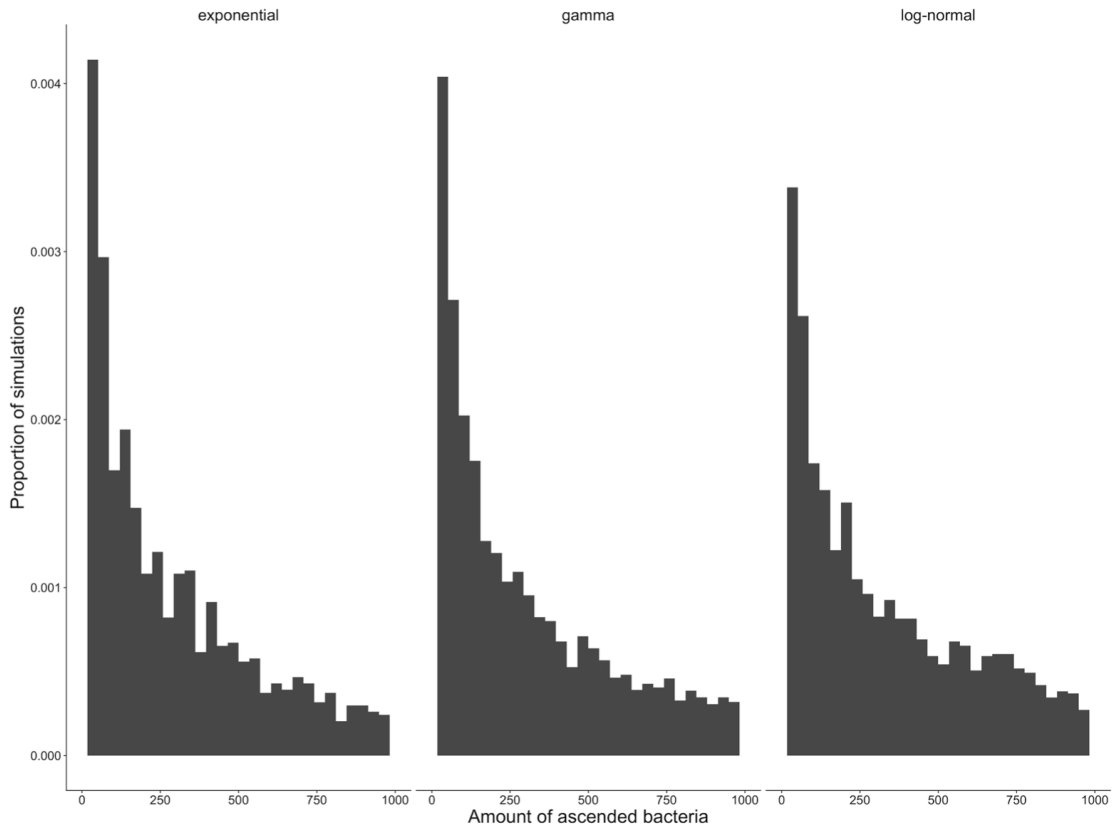

Figure 6.3: **A histogram of the amount of ascended bacteria under different distributional assumptions.** Each histogram represents 12,625 simulations from the model, where the underlying distribution for the rate parameters has been selected from an exponential distribution, a gamma distribution and a log-normal distribution. All distributions have identical means of 1/100.

We used probability distributions over key parameters of the model to adequately represent our uncertainty, and make our results less sensitive to assumptions. Some of the key distributions were assumed for this purpose, such as the rate parameter of the model. In order to test the sensitivity of our results to this assumption, we simulated the model where the rate parameters were drawn from exponential, gamma and log-normal distributions, with equivalent means. The results were compared for the key result of the amount of ascended bacteria (Figure 6.3).

6.3 Results

6.3.1 Modelling permits variation in ascension dynamics

Our modelling abstracts the ascension of *Chlamydia trachomatis* into a set of stochastic dynamics, that allows us to elucidate the moderating factors that might determine whether or not an infection ascends into the upper reproductive tract in women.

As outlined in the methods the factors include extracellular bacteria in the cervical epithelium infect cells which later burst and produce new amounts of bacteria. Immune responses, and in particular the presence of neutrophils, which eject material into the cervical lumen, and uterine peristalsis and the direction. Figure 6.4 indicates the factors that potentially determine chlamydial ascension. We evaluated three realisations of the same stochastic process with distinct parameter values, distinct initial conditions for infectious load and the same initial conditions for time until cervico-fundal peristaltic contractions and the initial amounts of infected cells and ejected materials to show the major outcomes of the model (A, B, C, refer to the same plots in the figure) indicating the proportions of ascensions influenced by each set of parameters. In scenario A, the input is a high infectious load. In this scenario the infection grows rapidly over time, however insufficient material is ejected from the epithelium and the infection will not ascend (as indicated by the lack of ejected EBs in pale blue in 3A). In scenario B, an output of the model where the input was a low infectious load, the infection is ejected at two distinct time points. However, this leaves insufficient material in the epithelium for the infection to replicate, and so the infection dies out there. Since the infection cannot replicate in the cervical canal, it to dies out well before peristalsis in the upwards direction begins. In the final example, scenario C, the infection is initiated with an even higher infectious load, and in this output from the model the infection survived in the epithelium whilst sufficient amounts of material are ejected (pale blue) for the infection to be present at the point where material can be transported into the upper genital tract.

Using the model we can examine the role of both burst rates (release of

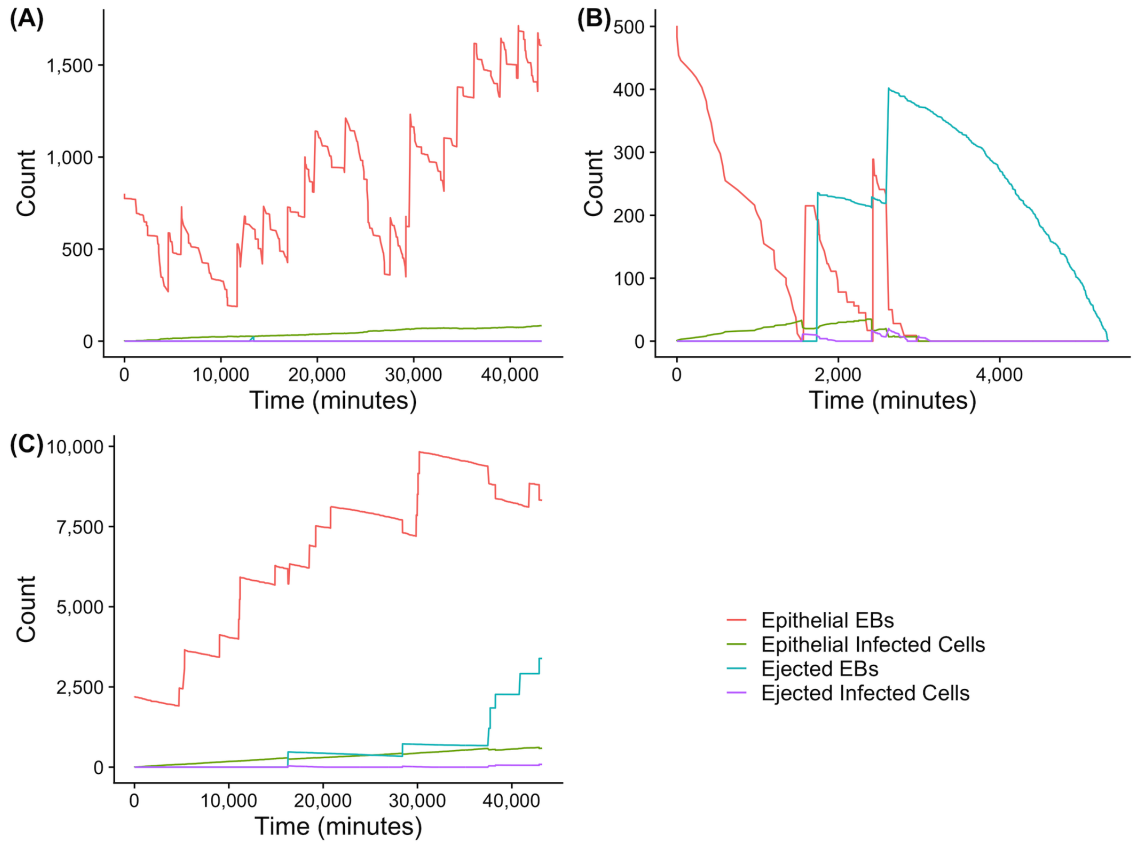

Figure 6.4: **A stochastic model to identify importance of governing factors in the model of ascension.** Graphical representation of the counts for each status of elementary bodies and cells in the model under three distinct scenarios over time. The figure shows the outputs of simulated sample paths of the model where factors have been modified. Each figure is a demonstration showing selected representations from the 252, 500 simulations that were conducted. (A) An example output from the model that starter with a load in the range of 801 EBs. (B) An example output of the model starting with a low initial load of 501 EBs. (C) An output that commended with a higher load than A and B of 2201 EBs. Each figure shows a graphical representation of count of each cell/chlamydia form represented on the y axis with the actual counts of *Chlamydia trachomatis* in the different forms indicated by the coloured lines (see figure legend bottom right), and time in minutes (x axis).

infectious *Chlamydia trachomatis* from cells) and inclusion formation rate (or our measure of establishing new inclusion vacuoles in new cells). This is based on the assumption that an increase in the frequency of bursts, would increase the number of new bacteria circulating in the epithelium. This is somewhat dependent on infectious dose, as higher infecting doses would have an increased frequency of bursts, but this factor reflects the infection across the course of time more generally. Increasing burst frequency does not appear to independently moderate the chances of ascension (Figure 6.5(A)). However, increasing the rate at which extracellular bacteria infect target cells does appear to increase the chances of ascension (Figure 6.5(B)). The chances of ascension when this rate was set at 0.001 infected cells per minute was estimated at 15.3%, 34.6% when the rate was 0.005 and 55.1 for a rate of 0.05.

Next, we examined simulations of the model to assess how frequently ascension occurs and the amount of bacteria that ascend. As Figure 5 demonstrates, even in infections where the bacteria ascend, there is a long tail in the distribution of how much bacteria will reach the upper genital tract. In our simulations, the median number of bacteria that ascended per infection was 0, with an interquartile range of 0 to 127. The top 25% of infections had between 127 and 59,090 bacteria ascended. This exercise gives an estimate of 36% of infections that ascend, but only 9% where more than 1000 bacteria managed to ascend, and 0.1% where more than 10,000 bacteria ascended.

6.3.2 Initial infectious load and timing relative to the menstrual cycle appear to moderate ascension

Overall, initial infectious load was influential on ascension events. Specifically, in this model we found initial loads in the range of 1000-3500 were most likely to ascend, with an apparent decrease in ascension found for higher loads (Figure 6.7(A)). The timing of initial infection relative to the onset of peristaltic contractions in the cervico-fundal direction (which is moderated by the menstrual cycle) had a much more marked impact on the simulated chances of ascension (Figure 6.7(B)). For example, infections that were initiated 30 hours prior to the onset of upwards peristaltic movement had a 7.8% of ascension, whereas infections initiated

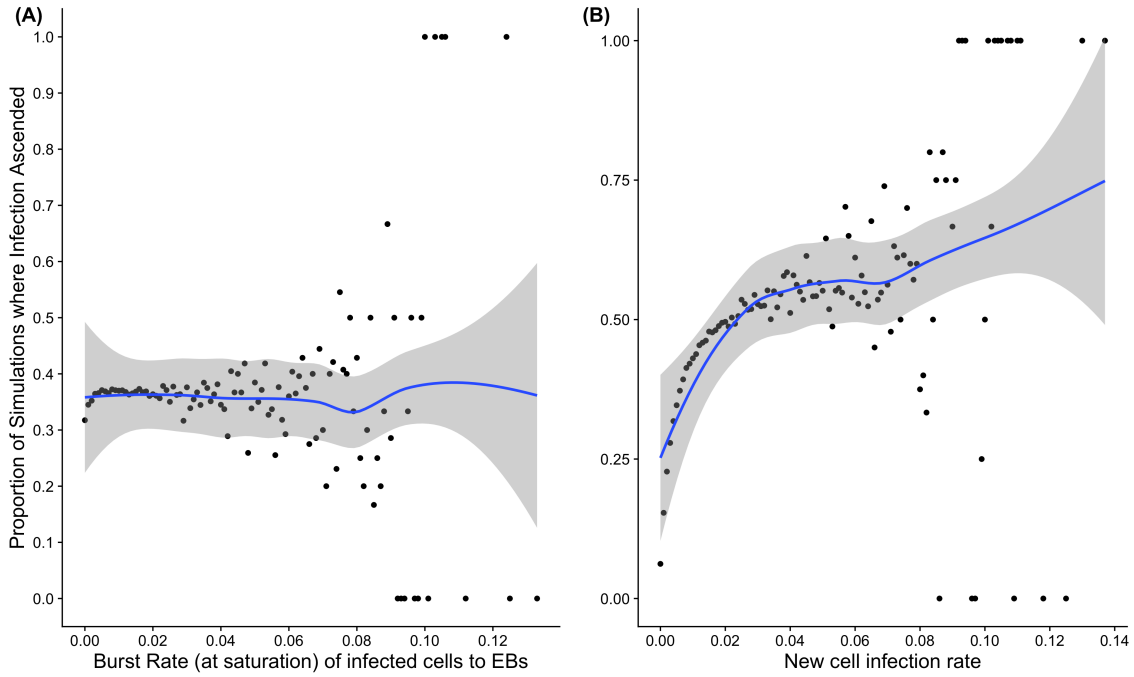

Figure 6.5: **Graphical representation of the proportion of ascension events in the simulations, as influenced by burst rate (A) and new cell infection rate (B).** (A) The relationship between burst rate (in cells per minute) and the simulated chances of ascension. (B) The relationship between new cell infection rate (cells per minute). The y axis shows the proportion of simulations where infections ascended (out of 252, 500 simulations conducted), and x axis shows the different parameters values of burst rate and new cell infection rate). Each dot represents simulations at a particular parameter value. The blue line represents a loess fit of ascension proportion against each parameter value, where shading is the 95% confidence interval of the fit.

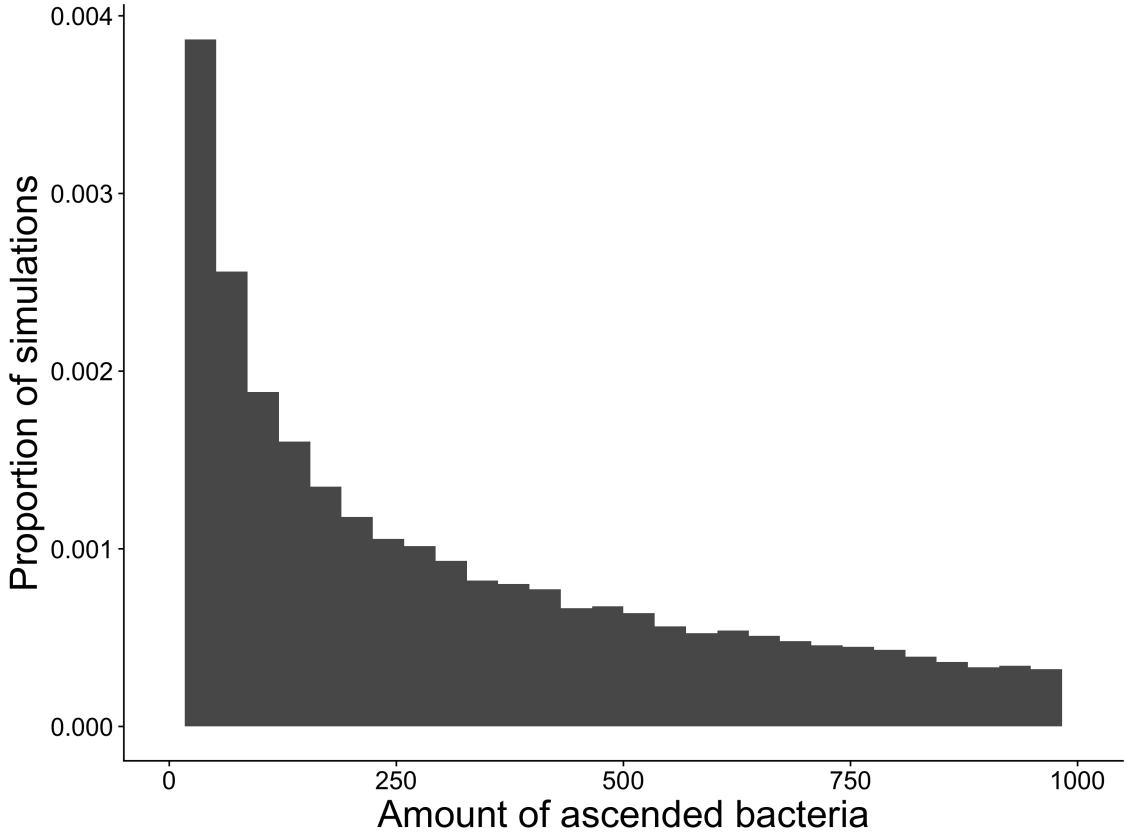

Figure 6.6: **Histogram of the proportion of simulations leading to ascension of bacteria, and the amount of bacteria predicted by the model to ascend.** The amount of bacteria predicted to ascend (x axis) and the proportions of each simulation that resulted in ascension (y axis) from 252, 500 simulations of the model.

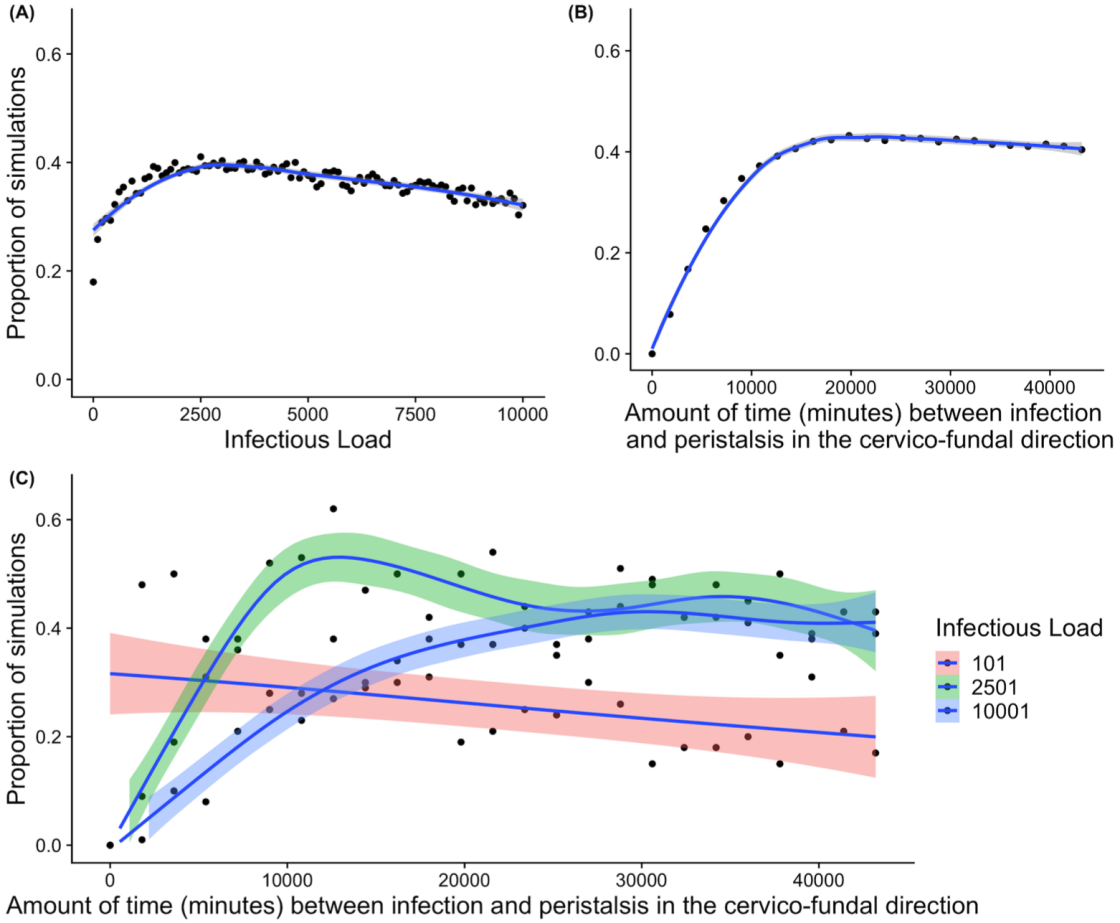

Figure 6.7: **Graphical summary of the proportions of simulations where ascension occurred and role of load (A), peristalsis (B), and both (C).** (A) Simulations analysing the impact of infectious load (the count of initial bacteria) on the the chances of ascension. (B) The time (in minutes) prior to the onset of peristaltic contractions in the cervico-fundal direction and chances of ascension. (C) The relationship between infectious load, timing of infection with respect to peristaltic contractions, and the chances of ascension. Each figure shows the proportion of simulations from the model where the infection ascended (proportion of representations from 252, 500 simulations) on the y axis, where the x axis represents the factor being modified (A). infectious load, (B). peristalsis, (C). peristalsis and grouped by infectious load). The blue lines represent a loess regression against the proportion of simulations resulting in ascension, and the shaded regions are 95% confidence intervals of the regression. Dots represent simulations at a particular parameter value.

10 days or more prior had a 42% of ascension. The relationship between infectious load and timing was also found to be critical in influencing ascension. Infections with smaller infectious loads experience linearly decreasing chances of ascension as the menstrual cycle progresses, so an infection that occurs later in the cycle may have reduced chances of ascending. However, the opposite applies to high infectious loads, where infections have a higher chance of ascending as the menstrual cycle progresses closer to peristaltic movement in the cervico-fundal direction (Figure 6.7(C)). Although a high initial load of 10001 had a broadly constant chance of ascension between 39-43%, for infections that were initiated 22 or more days prior to upwards peristaltic movement.

The relationship between infectious load and the menstrual cycle further complicates matters. Infections with smaller infectious loads experience linearly decreasing chances of ascension as the menstrual cycle progresses, so an infection that occurs later in the cycle may have reduced chances of ascending. However, the opposite applies to high infectious loads, where infections have a higher chance of ascending as the menstrual cycle progresses closer to peristaltic movement in the cervico-fundal direction.

6.3.3 Ascension chances are dependent on the neutrophil response, but appear independent of the rest of immune system response

Our simulation experiments show that the chances of ascension are broadly independent of the immune systems response (defined as overall scale of response, Figure 6.8), indicated by death rate of infectious material (extracellular EBs and infected cells) and moderated by the global immune response (separated into innate and adaptive systems). It is important to note that these phenomena represent the relative strength and weaknesses of the overall immune system and its responses between hosts. The model predicts that increasing or decreasing the strength of the host immune response, that determines the removal of EBs and infected cells, would have no impact on the chances of a particular infection ascending. The outcomes of the model under these parameters was that death of infected cells (Figure 6.8(B),

(D)) at a higher rate (rather than EBs, Figure 6.8(A), (C)) was more likely to reduce the proportion of infections that ascended in the simulations (although the outcomes of the model at high death rates were highly variable 0-100% of ascensions).

Infections that instigate higher neutrophil responses are more likely to be one that ascend (Figure 6.9). The simulated probability of ascension was 15.8% for a rate of 0.001 (per minute), and 42.5% for a simulated rate of 0.01. The exact nature of this relationship depends on the stage of the menstrual cycle in which infection occurs. It is an assumption of the model that the neutrophil response and other features of the innate and adaptive immune response are related by their density-dependence. That is, the ejection rate of cells due to the neutrophil response is higher when the removal rate of infected cells due to the innate and adaptive response is higher, as both respond to the overall density of infected cells.

6.4 Discussion

In this study we have applied a modelling approach to evaluate factors that influence chlamydial ascension in women. The impact of infectious burden on the likelihood of ascension shown as realisations in several iterations of the model presented here is not inconsistent with what might be expected from existing biological data in mouse and humans. It has been proposed that a strong adaptive immune system response may be sufficient to clear the infection from the cervix before ascension can occur [6]. Our modelling results indicate that ascension occurs independently of the strength of the immune system response, suggesting that the progression to pathological outcomes in the upper reproductive tract is not so straightforward. However, the presence of a neutrophil response has been previously associated with *Chlamydia trachomatis* and the development of pathology. There are a variety of studies that support the presence of neutrophils with Chlamydia and endometritis [31, 26]. Implicating the neutrophil responses in the movement of chlamydia, in spite of its role in controlling infections, and thus the findings here from our modelling are consistent with the literature on neutrophils. Our model assumes that the neutrophil response is critical in epithelium to the cervical canal, and so the neutrophil response plays a different role in modulating ascension to

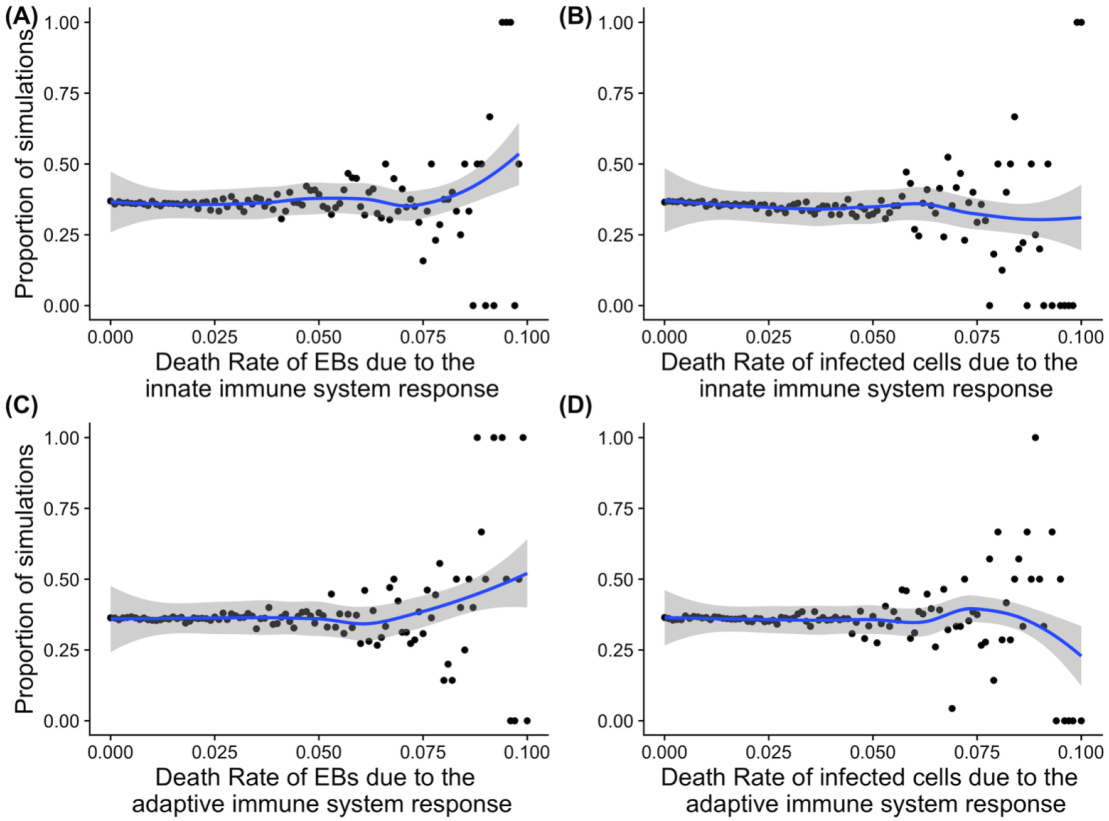

Figure 6.8: **Graphical representation of the impact of the immune response and subsequent chlamydial death rate on the proportion of simulations that result in ascension.** The proportion of simulations where the infection ascends was computed (total 252,500 simulations represented in each figure). Each figure shows the proportion of simulations where the infection ascended on the y axis and on the x axis the death rate of EBs and infected cells due to different components of the immune system (represented as broad scales of response). (A) Impact of EB death rate due to the innate immune system on the proportion ascending, (B) Impact of infected cells death rate due to the innate immune system on the proportion ascending, (C) Impact of EB death rate due to the adaptive immune system on the proportion ascending, and (D) Impact of infected cells death rate due to the adaptive immune system on the proportion ascending. The solid blue line indicates a loess fit to the proportion of simulations the ascended and the shaded area represents 95% Confidence Intervals of the loess regression. Dots represent simulations at a particular parameter value from a total of 252, 500.

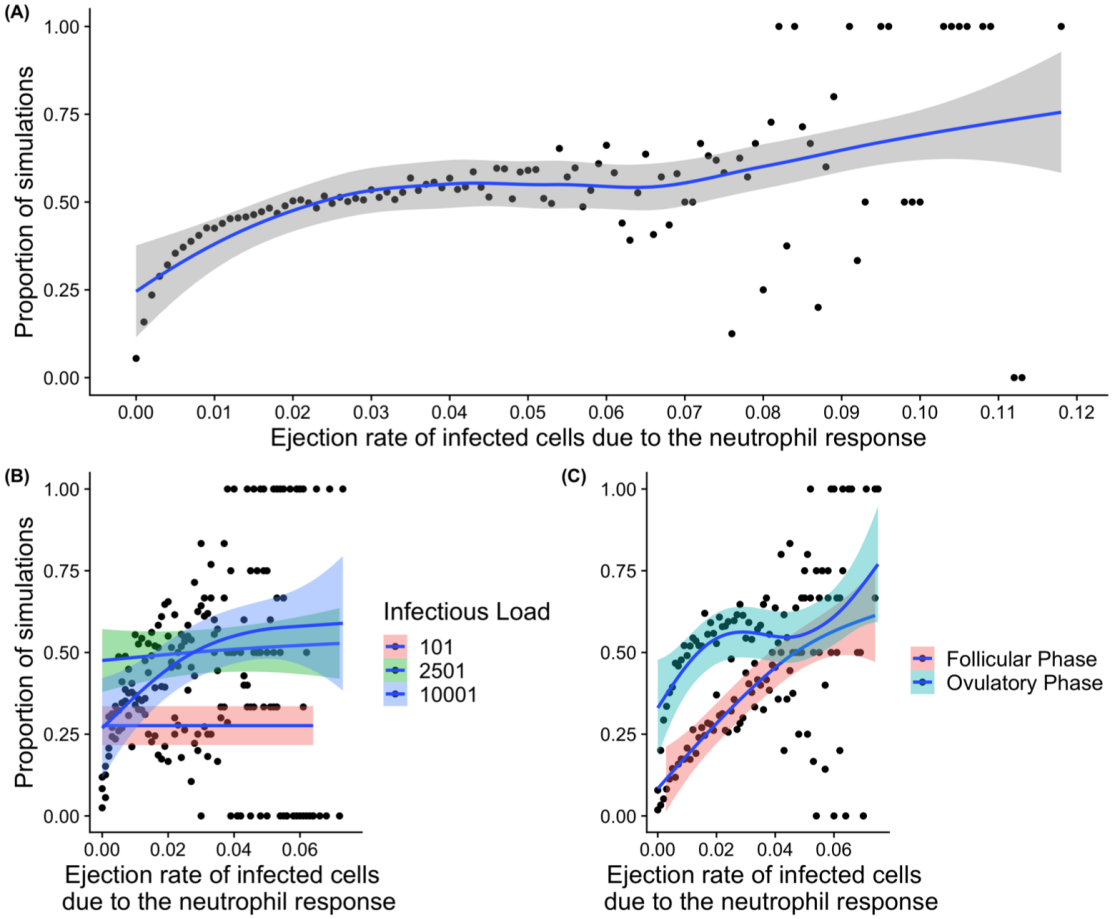

Figure 6.9: **Graphical representation of the proportion of infections that ascend in the simulations from the model and the role of the neutrophil response.** The figure shows the proportions of simulations (total of 252, 200 conducted for each instance) where infection was predicted to ascend (y axis). The ejection rate (movement to the cervix from the epithelium) of infected cells related to the neutrophil response i(x axis). (A) Ascension and the ejection rate due to the neutrophil response. (B) Impact of initial infectious load (pink represents a low infectious load of 101, green a moderate infectious load of 2501, and blue 10,001). (C) Time of infection relative to the menstrual cycle (follicular phase represented by pink, and ovulatory phase represented by blue). In each graph the solid blue line indicates a loess fit to proportion of simulations resulting in ascension, where the shaded area shows is the 95% confidence interval. Each dots are represents simulations at a particular parameter value, from a total of 252,500.

that observed from the global innate immune system, in regards to the relationship between high and low immune response and ascension [17]. It is not clear from the available literature if the neutrophil response varies independently between hosts, or if it occurs as a function of the amount of *Chlamydia trachomatis* present at a particular site in the epithelium (or both). From the perspective of the pathogen, one could view the evolution of *Chlamydia trachomatis* alongside the neutrophil response as a process that has selected for these adaptations to exploit this natural defence mechanism to ensure pathogen survival. The evidence we provide here suggests that an influx of neutrophils may be a useful predictor of an ascending infection.

A key assumption of this modelling exercise is that uterine peristalsis changes continuously with the menstrual cycle. For an infection to ascend, it requires exact timing with the menstrual cycle in order to utilise the peristaltic activity of the uterus. However, women with dysfunctions of the peristaltic mechanisms, such as hyperperistalsis, may experience peristaltic activity at three times the rate of women without these dysfunctions [13]. Further, there is some evidence that women with high uterine peristalsis have reduced pregnancy rates due to the presence of intramural fibroids [30]. Although speculation, it is plausible the women with increased peristaltic activity due to uterine fibroids would be at increased risk of ascending infections.

This interaction between the host's defence mechanism and the infection must be time-dependent for an infection to successfully ascend according to the outcomes shown here. As migration from the initial site of infection during the earlier phases of the menstrual cycle will more likely spread the infection too thin, with insufficient material left in the epithelium in order to ensure to short term survival of the infection and to ensure adequate infectious material for successful transmission to new hosts. Our simulations indicate that both the infecting load and the menstrual cycle moderate ascension, and that this moderation occurs in a dynamic and dependent way. There is limited biological data to compare with the proportions of infections that ascend that we have predicted here in the model. Ascension of chlamydia is a necessary, but not sufficient condition for pathology

(such as PID) to develop, and we know that PID occurs in approximately 9.5% of chlamydial infected women [19]. In a majority of our simulations where ascension did occur, the amount of material that ascended was relatively small (although 9% of ascensions in one model had greater than 1000 organisms, Figure 6.6). A case-control study of women with positive (cases) and negative (controls) tests for chlamydia, gonorrhoea or bacterial vaginosis, found the women in the follicular phase of their menstrual cycle had an adjusted odds ratio of 4.5 (95% CI (1.6 - 12.4) of experiencing upper genital tract infection [11]. Although the timing of biopsy in this clinical evidence is not necessarily equivalent to the timing of ascension, this study supports the findings we present that menstrual cycle moderates ascension through peristalsis. Our modelling finds that infections that are initiated in this stage of the menstrual cycle are least at risk of ascending.

The limitations of this work should be noted. This exercise characterises the process through simulated realisations of the model, and it is not calibrated or fit against biological datasets. However, an explicit mathematical representation of the dynamics of ascension should assist with further analysis of empirical outcomes. For example, although a number of parameters in this model (including the outcome) are difficult or impossible to measure, they provide useful information that could be incorporated into a prior distribution for a Bayesian analysis of some available empirical data. The scope of this work would be large, as there may be many different likelihoods for the dynamics we describe here based on the available data (such as immunological time series data, pathology data such as PID outcomes and infectious load data). In the next stage of research, a simulation study where data is generated according to the model described in this chapter, and then fitted with a statistical model would provide insight into the methods required to apply this model to empirical data.

Additionally, as with all modelling, our results rest on a set of assumptions that can potentially be modified to realise different outcomes. This cannot be avoided, although we have been as rigorous as possible in simulating over a broad range of parameter values and including stochastic terms to ensure the dynamic nature of the observations is encompassed. Unlike previous mathematical models of

chlamydia [15, 29, 28, 22], the model presented in this study incorporates stochastic elements that samples across a broad range of parameter values, which influences realised outcomes of the model and provides some robustness to the results presented [3]. Also, we do not continue the simulations beyond the point of upwards peristalsis within the stage of the cycle, meaning the longer term implications for the longer time course of the infection have not been considered in the model.

A significant component of chlamydial function not included in the model is the differing strategies that *Chlamydia trachomatis* uses to exit the cell. The mechanisms that govern lysis and extrusion cellular release pathways are distinct and may vary in the rate at which they occur and the amount of *Chlamydia trachomatis* released from the cell [10]. Extrusions of chlamydial inclusions have the ability to utilise macrophages as intracellular survival niches [32]. Chlamydia may then be able to exploit macrophages as a vehicle of dissemination within a host. The macrophage response has implications for other bacterial pathogens, in particular Neisseria gonorrhoeae, which has the ability to inhibit apoptosis and use macrophages as a replicative survival niche [5, 8].

In conclusion, we have proposed a process for the ascension of chlamydial infections, by considering the available clinical evidence and experimental observations of the pathogen and combining these observations into a stochastic model that allows us to understand the influential factors of ascension. The dynamics of the process are complex, nonlinear, and stochastic, so we model this in order to reason about the moderating factors. From our simulation results, we propose that relative cervical neutrophil responses and infections burden could be useful as early indicators of the likelihood of pathology. We should further examine neutrophil levels as predictive markers for pathology development of women at-risk of exposure to *Chlamydia trachomatis*. Further we propose that the model could be more broadly applied to other STI to identify relevant factors for ascension and pathology risk in the cases of those pathogens.

References

[1] Yasser M. AbdelRahman and Robert J. Belland. "The chlamydial developmental cycle". In: *FEMS Microbiology Reviews* 29.5 (2005), pp. 949–959. URL: https://doi.org/10.1016/j.femsre.2005.03.002.

[2] N Barteneva et al. "Role of neutrophils in controlling early stages of a Chlamydia trachomatis infection". In: *Infection and Immunity* 64.11 (1996), pp. 4830–4833. URL: https://iai.asm.org/content/iai/64/11/4830.full.pdf.

[3] Torrington Callan et al. "A retrospective cohort study examining STI testing and perinatal records demonstrates reproductive health burden of chlamydia and gonorrhea". In: *Pathogens and Disease* 78.6 (2020). ISSN: 2049-632X. DOI: 10.1093/femspd/ftaa052. URL: https://doi.org/10.1093/femspd/ftaa052.

[4] Alison J. Carey et al. "Effects of inoculating dose on the kinetics of Chlamydia muridarum genital infection in female mice". In: *Immunology and Cell Biology* 87.4 (2009), pp. 337–343. ISSN: 0818-9641, 0818-9641. DOI: http://dx.doi.org/10.1038/icb.2009.3.

[5] Alice Château and H. Steven Seifert. "Neisseria gonorrhoeae survives within and modulates apoptosis and inflammatory cytokine production of human macrophages". In: *Cellular microbiology* 18.4 (2016), pp. 546–560. ISSN: 1462-5822 1462-5814. DOI: 10.1111/cmi.12529. URL: https://pubmed.ncbi.nlm.nih.gov/26426083%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5240846/.

[6] Toni Darville and Thomas J. Hiltke. "Pathogenesis of genital tract disease due to Chlamydia trachomatis". In: *The Journal of infectious diseases* 201 Suppl 2.Suppl 2 (2010), S114–S125. ISSN: 1537-6613 0022-1899. DOI: 10.1086/652397. URL: https://pubmed.ncbi.nlm.nih.gov/20524234%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3150527/.

[7] Bethan Davies et al. "Risk of reproductive complications following chlamydia testing: a population-based retrospective cohort study in Denmark". In: *The Lancet Infectious Diseases* 16.9 (2016), pp. 1057–1064. ISSN: 1473-3099. DOI:

10.1016/S1473-3099(16)30092-5. URL: https://doi.org/10.1016/S1473-3099(16)30092-5.

[8] Alejandro Escobar, Paula I. Rodas, and Claudio Acuña-Castillo. "Macrophage–Neisseria gonorrhoeae Interactions: A Better Understanding of Pathogen Mechanisms of Immunomodulation". In: *Frontiers in Immunology* 9.3044 (2018). ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.03044. URL: https://www.frontiersin.org/article/10.3389/fimmu.2018.03044.

[9] Catherine L. Haggerty et al. "Risk of Sequelae after Chlamydia trachomatis Genital Infection in Women". In: *The Journal of Infectious Diseases* 201 (2010), S134–S155. URL: https://doi.org/10.1086/652395.

[10] Kevin Hybiske and Richard S. Stephens. "Mechanisms of host cell exit by the intracellular bacterium ¡em¿Chlamydia¡/em¿". In: *Proceedings of the National Academy of Sciences* 104.27 (2007), pp. 11430–11435. DOI: 10.1073/pnas.0703218104. URL: https://www.pnas.org/content/pnas/104/27/11430.full.pdf.

[11] Abner P. Korn et al. "Risk factors for plasma cell endometritis among women with cervical Neisseria gonorrhoeae, cervical Chlamydia trachomatis, or bacterial vaginosis". In: *American Journal of Obstetrics and Gynecology* 178.5 (1998), pp. 987–990. ISSN: 0002-9378. DOI: https://doi.org/10.1016/S0002-9378(98)70536-8. URL: http://www.sciencedirect.com/science/article/pii/S0002937898705368.

[12] Georg Kunz and Gerhard Leyendecker. "Uterine peristaltic activity during the menstrual cycle: characterization, regulation, function and dysfunction". In: *Reproductive biomedicine online* 4 Suppl 3 (2002), pp. 5–9. URL: https://www-lib-uts-edu-au.ezproxy.lib.uts.edu.au/goto?url=http://search.ebscohost.com.ezproxy.lib.uts.edu.au/login.aspx?direct=true&db=mnh&AN=12470555&site=ehost-live.

[13] Gerhard Leyendecker et al. "Uterine peristalsis and dysperistalsis as dysfunction of the mechanism of rapid sperm transport in patients with endometriosis and infertility". In: *Human reproduction (Oxford, England)* 11 (1996), pp. 1542–51.

[14] N. Low et al. "Incidence of severe reproductive tract complications associated with diagnosed genital chlamydial infection: the Uppsala Women's Cohort Study". In: *Sex Transm Infect* 82.3 (2006), pp. 212–218.

[15] Dann G. Mallet et al. "A Mathematical Model of Chlamydial Infection Incorporating Movement of Chlamydial Particles". In: *Bulletin of Mathematical Biology* 75.11 (2013), pp. 2257–2270. ISSN: 1522-9602. DOI: 10.1007/s11538-013-9891-9. URL: https://doi.org/10.1007/s11538-013-9891-9.

[16] Dann G. Mallet et al. "Chlamydial infection and spatial ascension of the female genital tract: a novel hybrid cellular automata and continuum mathematical model". In: *FEMS Immunology And Medical Microbiology* 57.2 (2009), pp. 173–182. ISSN: 1574-695X. DOI: 10.1111/j.1574-695X.2009.00596.x. URL: https://www.lib.uts.edu.au/goto?url=http://search.ebscohost.com/login.aspx?direct=true&db=mnh&AN=19735471&site=ehost-live.

[17] Heather K. Maxion et al. "The Infecting Dose of Chlamydia muridarum Modulates the Innate Immune Response and Ascending Infection". In: *Infection and Immunity* 72.11 (2004), pp. 6330–6340. DOI: 10.1128/iai.72.11.6330-6340.2004. URL: https://iai.asm.org/content/iai/72/11/6330.full.pdf.

[18] S. Menon et al. "Human and Pathogen Factors Associated with Chlamydia trachomatis-Related Infertility in Women". In: *Clinical Microbiology Reviews* 28.4 (2015), pp. 969–985. URL: https://cmr.asm.org/content/28/4/969.

[19] Pippa Oakeshott et al. "Randomised controlled trial of screening for Chlamydia trachomatis to prevent pelvic inflammatory disease: the POPI (prevention of pelvic infection) trial". In: *BMJ* 340 (2010). URL: https://www.bmj.com/content/340/bmj.c1642.

[20] John E. Pearson, Paul Krapivsky, and Alan S. Perelson. "Stochastic Theory of Early Viral Infection: Continuous versus Burst Production of Virions". In: *PLoS Computational Biology* 7.2 (2011), e1001058. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1001058.

[21] Roger G. Rank et al. “Chlamydiae and polymorphonuclear leukocytes: unlikely allies in the spread of chlamydial infection”. In: *FEMS Immunol Med Microbiol* 54.1 (2008), pp. 104–113.

[22] O. Sharomi and AB. Gumel. “Mathematical study of in-host dynamics of Chlamydia trachomatis”. In: *Journal of Theoretical Biology* 77.2 (2012), pp. 109–139.

[23] C. Stacey et al. “Chlamydia trachomatis in the fallopian tubes of women without laparoscopic evidence of salpingitis”. In: *International Journal of Gynecology and Obstetrics* 35.3 (1991), p. 287. URL: http://www.sciencedirect.com/science/article/pii/002072929190324X.

[24] Lenka A. Vodstrcil et al. “The Epidemiology of Chlamydia trachomatis Organism Load During Genital Infection: A Systematic Review”. In: *The Journal of Infectious Diseases* 211.10 (2014), pp. 1628–1645. ISSN: 0022-1899. DOI: 10.1093/infdis/jiu670. URL: https://doi.org/10.1093/infdis/jiu670.

[25] Lars Westrom et al. “Pelvic Inflammatory Disease and Fertility”. In: *Sexually Transmitted Diseases* 19.4 (1992), pp. 185–192.

[26] Harold C. Wiesenfeld et al. “Association between Elevated Neutrophil Defensin Levels and Endometritis”. In: *The Journal of Infectious Diseases* 186.6 (2002), pp. 792–797. URL: https://doi.org/10.1086/342417.

[27] D. P. Wilson. “Mathematical modelling of Chlamydia”. In: *Proc. of 11th Computational Techniques and Applications Conference CTAC-2003*. Ed. by Jagoda Crawford and A. J. Roberts. Vol. 45, pp. C201–C214.

[28] D.P. Wilson and D.L.S. McElwain. “A model of neutralization of Chlamydia trachomatis based on antibody and host cell aggregation on the elementary body surface”. In: *Journal of Theoretical Biology* 226.3 (2004), pp. 321–330. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2003.09.010.

[29] David Wilson. “Mathematical Modelling of Chlamydia”. In: *The ANZIAM Journal* 45 (Apr. 2004). DOI: 10.21914/anziamj.v45i0.883.

[30] O. Yoshino et al. “Decreased pregnancy rate is linked to abnormal uterine peristalsis caused by intramural fibroids”. In: *Human Reproduction* 25.10 (2010), pp. 2475–2479. URL: https://doi.org/10.1093/humrep/deq222.

[31] Mark H. Yudin et al. “Vaginal polymorphonuclear leukocytes and bacterial vaginosis as markers for histologic endometritis among women without symptoms of pelvic inflammatory disease”. In: *American Journal of Obstetrics and Gynecology* 188.2 (2003), pp. 318–323. URL: http://www.sciencedirect.com/science/article/pii/S000293780271435X.

[32] M. Zuck et al. “Extrusions are phagocytosed and promote Chlamydia survival within macrophages”. In: *Cell Microbiol* 19.4 (2017). ISSN: 1462-5814. DOI: 10.1111/cmi.12683.

7

Conclusions and Future Research

In this chapter, we summarise our findings from each chapter this thesis, and discuss the future avenues of research that builds on the results we have so far. The idea of using both mathematical and statistical modelling approaches is discussed. We then summarise the conclusions of the results found in this thesis.

7.1 Summary of findings

In Chapter 2, we used a retrospective data linkage cohort study to investigate the reproductive burden of testing and testing diagnosis for Chlamydia and/or Gonorrhea. Using statistical modelling techniques, it was found that testing status, independent of the test result, was the biggest predictor of pregnancy as substantiated by a birth record in the study outcome window. There were differences in pregnancy outcomes between women with all negative tests and women with single positive tests, and these differences were highly age-specific (for example, women with all negative tests were least likely to have a pregnancy at ages 20 and 30, and women with a single positive test were least likely when aged 25). These results confirmed the substantial reproductive burden associated with testing for these two pathogens, but simultaneously introduced a more complex view of the processes involved. Previous literature measuring Pelvic Inflammatory Disease (PID) as an outcome found clear differences between women with no testing history, all negative tests and positive tests. Our results in chapter 2 found a more complicated relation-

ship between testing, infection and reproductive burden as measured by pregnancy, which suggests the presence of unmeasured confounders.

In Chapter 3, we investigated a relevant confounder to our data, that of Indigenous status. Indigenous women as a population have younger ages compared to the non-Indigenous population at which they become pregnant, and in our data were more likely to have multiple positive tests. Indigenous status was measured differently at different stages of collection of the data used in our study, and so measurement error corrections were applied using a multiple imputation procedure. This correction allowed the highly confounded effects of Indigenous status and multiple positive tests to be separated somewhat. After applying the imputation process, we also found that there was no significant difference in the odds of pregnancy between women with all negative tests and women with single positive tests.

In Chapter 4, we investigate the relationship between the usage of ART and testing history on the perinatal outcome of birth weights. In order to fully capture the distribution of birth weights, a mathematical model of birth weight as a function of gestational age was derived. The mathematical model was then developed into a statistical one with the addition of parameter-level covariates. Using this approach, we found that women with usage of ART and women with a history of testing had distinct patterns of birth weights over comparable gestational periods, implying that the nature of infertility for women with a history of testing was distinct from that experienced by women accessing ART.

In Chapter 5, we set up a basic stochastic within-host model of chlamydia using a framework of branching processes. This extended previous deterministic within-host mathematical models on chlamydial dynamics. We used simulated realisations from the model to investigate global properties of a chlamydial infection, including time to clearance and overall infectious burden. We then ran a series of sensitivity analyses of the model using idealised representations of the immune system to investigate which immune system response would be most effective at clearing an infection.

In Chapter 6, we used the model developed in Chapter 5 to investigate the

moderating factors of ascension, that allows for an infection to ascend beyond the cervix to the upper genital tract. As ascension is difficult to empirically measure, and the moderating factors interact in complex and nonlinear ways, this hypothesis was amenable to investigation via a mathematical modelling approach. We found that the menstrual cycle, initial infectious load and features of the immune response were critical in determining which infection ascended, and so determined at set of markers for women potentially at-risk of developing pathology. This concept relates to the results in Chapters 2 and 3, which suggests that the reproductive burden of testing and infection vary by key demographic and biological features.

7.2 Future research

7.2.1 Longitudinal modelling of the reproductive burden of testing

Whilst our data used in Chapter 2 has significant limitations that prevents us from doing a full discrete-time based analysis in a longitudinal framework, the results presented in this chapter, particularly the discrete time analysis demonstrate the potential importance of considering the time-based effects. There is great future opportunity in gathering full longitudinal used to investigate more precise hypotheses about the relationship between testing, pathogens and reproductive burden. For example, as seen above the impact of testing may be more apparent in younger women or closer to the testing date.

A study of this format would allow for better precision of the risk associated with particular testing patterns. As our data have indicated, women with all negative tests may have lower chances of pregnancy, compared to women with a single positive test. However, this statement does not take into account the nature of negative tests for women with only a single positive one.

Longitudinally collected data that includes testing statues across time, with key demographic variables include (such as Age, SES and ethnicity) could be powerfully combined with immunological, microbiome and bacterial burden data, using a

combination of the models and methods described in this thesis. Statistical models and supporting methods that estimate the reproductive burden, such as the ones described in the earlier chapters, could be informed by simulations of an underlying mathematical model such as the ones in the later chapters of this thesis. This broad approach would combine knowledge from clinical and observational studies with experimental data to provide the a comprehensive and systematic view of the various influencing factors of pathogen response, the associated social dynamics around testing and reproductive burden.

7.2.2 Improving the basic chamydial model

The basic model of chlamydia introduced in Chapter 5 and utilised to investigate the ascension hypothesis in Chapter 6 could be improved in a number of ways to capture various phenomena that has been noted in the literature to impact chlamydial growth dynamics.

Our model does not explicitly track the biphasic developmental cycle of chlamydia, rather, it measures the outcome of the developmental cycle in terms of burst and/or lysis size. Other models have been developed to describe the conversion from EB to RB forms and vice versa, including some of the stochastic features that determine host death, and terminal conversion time [13, 4]. Of particular interest, these models provide a bio-theoretical basis for the observed features of the developmental cycle, that matches experimental data from three dimensional electron microscopy [7]. Utilising these replication models in a multiscale format, combined with the model proposed in this thesis, could provide a more robust understanding of the dynamics of chlamydial growth, that is grounded in a theoretical understanding of the pathogen.

As discussed in Chapter 5, Chlamydia has two distinct exit strategies from a cell, lysis and extrusions, which are mediated by distinct mechanisms [6]. The extrusion phenomena differs from lysis in terms of the amount and rate of exit, and in turn these features may reflect host and serovar specific factors. Extending the model to specifically include extrusions would give it greater biological realism and

clinical relevance. Further, extrusions influence the spatial spread of chlamydia by utilising macrophages as an intracellular survival niche [15]. An interesting piece of future research would be to compare the ascension model via the neutrophil response and uterine peristalsis proposed in Chapter 5, to the macrophage response suggested by the research on extrusions.

The immune response and the interaction with the chlamydial life cycle in Chapter 4 and Chapter 5 are under specified compared to the available research and data on immunology and chlamydia [1]. For example, in Chapter 4 we made broad assumptions about the interactions between components of the immune response and the specific impact on chlamydia, and in Chapter 5, we described the global innate and adaptive immune response and characterised how it might influence the dynamics of the model. However, we could include greater specificity on the immunology, and potentially combine with available longitudunal data, to improve the immune components of the model.

7.2.3 The relationship between the cervico-vaginal microbiome and ascension of Chlamydia

The vaginal microbiome of women of reproductive age has been characterised into five community state types (CST), four of which are dominated by species of Lactobacillus (I crispatus, II gasseri, III iners, V jensenii), and the fifth of which lacks a dominance of Lactobacillus and is instead defined by the presence of a diverse number of anaerobes [8].

Bidirectional relationships between the composition of the cervico-vaginal microbiome and susceptibility to Chlamydia trachomatis infections have been demonstrated by a number of epidemiological studies [10, 14, 12]. The bidirectionality of the relationships suggest a number of behavioural and biological factors (such as micronutrient availability [9]) modulate both the composition of the microbiome and the presence of Chlamydia in the lower genital tract, although the community state type of the host differentiates the risk of infection for an individual after adjustment of key behavioural confounders [10, 11, 5]. The available evidence suggests that the

microbiota-host interaction modulates protective effects against infection [3].

The reproductive health burden of Chlamydia trachomatis has been demonstrated through the association with infection and pelvic inflammatory disease (PID) and tubal factor infertility, and has been discussed further in Chapters 2 and 3. It has been proposed that the tubal damage associated with an infection occurs when the pathogen ascends the genital tract and instigates a pro-inflammatory response in the upper genital tract. The mathematical modelling in Chapter 5 has indicated that the probability of ascension is modulated by a number of key host and pathogen specific factors, including the immune response, reproductive cycle and the proliferation of the pathogen. A number of the key modulating factors of ascension are correlated with microbial composition of the cervix and vagina, such as the innate immune response [2].

In is therefore natural to ask, how might the relationship between the host, pathogen and the cervico-vaginal microbiome modulate ascension of Chlamydia. In particular, are there CSTs that are better are protecting against ascension, and therefore the pro-inflammatory consequences of infection? We might suspect that the answer is identical to those CSTs that are most protective of infection overall, but as the relationship between infection and ascension is complex, this may not necessarily be the case.

7.3 Discussion on combining modelling approaches

In the introduction to this thesis, a distinction between mathematical and statistical modelling approaches was described. Across the results presented, a number of methods have been applied in order to understand the reproductive burden of STI testing, how this varies between key demographics, and how host and pathogen factors may moderate the progression of infection to pathology and contribute to this reproductive burden.

Mathematical and statistical modelling are usually thought of as two unique and disparate practices. However, the two are very closely linked. A statistical model will have a mathematical representation as a core component, In fact the only part

that is not also present in a mathematical model is the inclusion of a likelihood that allows the model to be fit to a set of data (although techniques such as Approximate Bayesian Computation do not require specification of a likelihood). In contrast, a mathematical model will often be more considered and complex, as the practice often draws from multiple sources of literature in order to form a coherent model (this is the process employed to derive the model in Chapter 6).

In Chapter 4, we showed how using a derived mathematical model in order to understand a new set of data led to a deeper understanding to the process of gestation and birth weights and the potential impacts of ART usage and testing. This is a modelling process that can and perhaps should be more widely employed. It relies on a prior state of knowledge and an understanding of the problem being solved, rather than fallible model selection techniques such as forwards or backwards variable selection for regression models. A more approximate version of this was used to select variables for the models in Chapter 2 and 3. The process being studied was more complicated and less identifiable from the data, and so a richer mathematical model for pregnancy would be difficult to derive. However, principled variable selection and transformations (using splines and interactions), still approximates this approach.

The ideal would be to take a well-defined mathematical model such as the one developed in Chapters 5 and 6, and use this to interpret new sets of data. Although difficult to achieve in practice, aiming for this would allow for greater accountability in modelling exercises, and hopefully allow for deeper scientific understanding. It is an approach that is quite Bayesian in flavour. A prior state of knowledge that exists within the scientific literature, and in the collective minds of the scientists studying the process. The prior state of knowledge can be updated by viewing new empirical observations. These new empirical observations may be newer studies, or new datasets available for analysis, which is used to form a new state of knowledge.

Bayesian methods have been used extensively throughout this thesis, as the computational methods allow for complex and flexible models to be fit to the available data. Probability models can be specified and fit to data, and in turn usage

of Bayesian computational approaches allows for more complex probability models to be specified than were previously possible.

In this way, Bayesian inference is a unifying idea between statistical and mathematical modelling approaches. Under this framework, the distinction between the two refers to the data (likelihood) being used to update our beliefs. In most statistical approaches, data refers to experimental or observational data that contains replicates of the same process being studied. In most mathematical modelling approaches, the data refers to different studies and observations made within these. Previous studies, meta-analyses and population level observations can be added to the prior (more precisely, literally added to the log of the prior) in a univariate or multivariate way, which can be combined with marginal prior distributions of individual parameters that are then informed by experimental or observational data.

7.4 Conclusion

Overall, this thesis has utilised a collection of statistical and mathematical modelling approaches, to elucidate some of the key influential factors of the development of pathology and reproductive burden. Mathematical models are powerful at combining known features of the literature into a coherent model, and producing new results for further empirical investigation. However, they are difficult to validate against empirical evidence. Inclusion of stochastic terms in these models allow for the representation of intrinsic uncertainty. Statistical models are useful as tools for sophisticated data analysis, and easily allow for representation of parameter uncertainty. A synthesis of these two ideas would be a powerful way of problem solving, where principled information from the literature on a certain problem could be used to form the structure of a mathematical model with a strong mechanistic interpretation, and then fit to appropriate datasets for a more detailed statistical analysis.

References

[1] Robert C. Brunham and José Rey-Ladino. "Immunology of Chlamydia infection: implications for a Chlamydia trachomatis vaccine". In: *Nature Reviews Immunology* 5.2 (2005), pp. 149–161. ISSN: 1474-1741. DOI: 10.1038/nri1551. URL: https://doi.org/10.1038/nri1551.

[2] Sylvie Y. Doerflinger, Andrea L. Throop, and Melissa M. Herbst-Kralovetz. "Bacteria in the Vaginal Microbiome Alter the Innate Immune Response and Barrier Properties of the Human Vaginal Epithelia in a Species-Specific Manner". In: *The Journal of Infectious Diseases* 209.12 (2014), pp. 1989–1999. ISSN: 1537-6613. DOI: 10.1093/infdis/jiu004. URL: https://dx.doi.org/10.1093/infdis/jiu004.

[3] Vonetta L. Edwards et al. "The Cervicovaginal Microbiota-Host Interaction Modulates Chlamydia trachomatis Infection". In: *mBio* 10.4 (2019). ISSN: 2150-7511. DOI: 10.1128/mbio.01548-19. URL: https://dx.doi.org/10.1128/mbio.01548-19.

[4] German Enciso et al. "Stochastic Chlamydia Dynamics and Optimal Spread". In: *Bulletin of Mathematical Biology* 83.4 (2021). ISSN: 0092-8240. DOI: 10.1007/s11538-020-00846-4.

[5] Simone Filardo et al. "Diversity of Cervical Microbiota in Asymptomatic Chlamydia trachomatis Genital Infection: A Pilot Study". In: *Frontiers in Cellular and Infection Microbiology* 7 (2017). ISSN: 2235-2988. DOI: 10.3389/fcimb.2017.00321. URL: https://dx.doi.org/10.3389/fcimb.2017.00321.

[6] Kevin Hybiske and Richard S. Stephens. "Mechanisms of host cell exit by the intracellular bacterium ¡em¿Chlamydia¡/em¿". In: *Proceedings of the National Academy of Sciences* 104.27 (2007), pp. 11430–11435. DOI: 10.1073/pnas.0703218104. URL: https://www.pnas.org/content/pnas/104/27/11430.full.pdf.

[7] Jennifer K Lee et al. "Replication-dependent size reduction precedes differentiation in Chlamydia trachomatis." In: *Nat Commun* 9.1 (Jan. 2018), p. 45.

[8] J. Ravel et al. “Vaginal microbiome of reproductive-age women”. In: *Proceedings of the National Academy of Sciences* 108.Supplement$_1$ (2011), pp. 4680–4687. ISSN: 0027-8424. DOI: 10.1073/pnas.1002611107. URL: http://www.pnas.org/content/108/Supplement_1/4680.full.pdf.

[9] Susan Tuddenham et al. “Associations between dietary micronutrient intake and molecular-Bacterial Vaginosis”. In: *Reproductive Health* 16.1 (2019). ISSN: 1742-4755. DOI: 10.1186/s12978-019-0814-6. URL: https://dx.doi.org/10.1186/s12978-019-0814-6.

[10] Janneke H. H. M. Van De Wijgert. “The vaginal microbiome and sexually transmitted infections are interlinked: Consequences for treatment and prevention”. In: *PLOS Medicine* 14.12 (2017), e1002478. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002478. URL: https://dx.doi.org/10.1371/journal.pmed.1002478.

[11] Charlotte Van Der Veer et al. “The Cervicovaginal Microbiota in Women Notified for Chlamydia trachomatis Infection: A Case-Control Study at the Sexually Transmitted Infection Outpatient Clinic in Amsterdam, The Netherlands”. In: *Clinical Infectious Diseases* 64.1 (2017), pp. 24–31. ISSN: 1058-4838. DOI: 10.1093/cid/ciw586. URL: https://dx.doi.org/10.1093/cid/ciw586.

[12] Robin Van Houdt et al. “Lactobacillus iners-dominated vaginal microbiota is associated with increased susceptibility to Chlamydia trachomatis infection in Dutch women: a case–control study”. In: *Sexually Transmitted Infections* 94.2 (2018), pp. 117–123. ISSN: 1368-4973. DOI: 10.1136/sextrans-2017-053133. URL: https://dx.doi.org/10.1136/sextrans-2017-053133.

[13] Frederic Y M Wan and German A Enciso. “Optimal Proliferation and Differentiation of Chlamydia Trachomatis.” In: *Stud Appl Math* 139.1 (July 2017), pp. 129–178. DOI: 10.1111/sapm.12175.

[14] Harold C. Wiesenfeld et al. “Bacterial Vaginosis Is a Strong Predictor of Neisseria gonorrhoeae and Chlamydia trachomatis Infection”. In: *Clinical Infectious Diseases* 36.5 (2003), pp. 663–668. ISSN: 1058-4838. DOI: 10.1086/367658. URL: https://doi.org/10.1086/367658.

[15] M. Zuck et al. “Extrusions are phagocytosed and promote Chlamydia survival within macrophages”. In: *Cell Microbiol* 19.4 (2017). ISSN: 1462-5814. DOI: 10.1111/cmi.12683.