Integrating Data, Demography, and Dynamics to Inform Vaccination Policy: Measles and Rubella in a Changing World

Amy Kaye Winter

A Dissertation
Presented to the Faculty of Princeton University
in Candidacy for the Degree of Doctor of Philosophy

Recommended for Acceptance
by the Program in Population Studies
Adviser: C. Jessica E. Metcalf

November 2016
Abstract

In this dissertation, I explore public health questions related to the current status of measles and rubella incidence and immunity, and the control of measles and rubella infections. Great strides have been made in the reduction of measles and rubella infections via vaccination. However, measles remains a significant cause of morbidity (20 million annual cases) and mortality (115,000 annual deaths); and rubella infection, typically mild among childhood, among pregnant women can cause the birth of an infant with congenital rubella syndrome (CRS) (105,000 annual CRS cases).

I built on an existing age-structured mathematical model to evaluate important public health questions in the context of measles and rubella control efforts. Mathematical models are increasingly relied upon for vaccination program design because formally framing the core mechanisms associated with transmission dynamics is essential to capturing important and potentially devastating non-linear effects in the dynamics of these infections.

In chapter 1, I evaluated the potential of introducing rubella-containing vaccine in India to reduce the burden of CRS, and found that introduction of the rubella vaccine will likely result in cumulative decreases in CRS. However, the effect of rubella vaccine introduction on the transient annual incidence of CRS is highly sensitive to rubella’s basic reproductive number. We identified risk factors that can be used to highlight regions most at risk of transient increases in CRS burden post-rubella vaccine introduction.

In chapter 2, I explored the use of a nested serological survey within the fever-rash surveillance system in Madagascar to estimate measles population immunity. We found discrepancies between direct and indirect empirical estimates of population immunity by age. However, both estimates indicated that Madagascar is at risk of a large measles outbreak. We evaluated the potential of measles Supplementary Immunization Activities to reduce this risk.

In chapter 3, I analyzed the strengths and limitations of rubella IgG serological data to characterize rubella epidemiological parameters and CRS incidence. We laid out in detail the nuanced biases from analytic method and survey design, which can be used by public health officials to better in-
interpret past serological survey estimates, and to design, implement, and interpret future serological surveys.
Acknowledgements

I am thankful to the many individuals who have supported my research endeavors in global health, and who have assisted in the completion of this dissertation.

First, I thank my committee members, Jess Metcalf, Bryan Grenfell, and James Trussell. I have been so fortunate to have a committee who truly support my research and career, and from whom there is so much to learn. To me, Jess is the epitome of how to be a confident, intelligent, and kind academic. She fluidly melds disciplines and personalities to bring the best out of the individuals she works with. Her work is thoughtful, thorough, and innovative. If one day I approximate to this, it will be largely due to the unwavering support, confidence, and insight she has provided me. I am privileged to work with her and hope that I make her proud. I am so grateful to Bryan who decided that I would be a good fit with Jess. Bryan has an infectious passion for work on disease dynamics that I am happy to have caught. James Trussell, who seemingly no matter what time zone, answers my emails within seconds, has been a strong supporter of my graduate student research, and has been flexible to fill any needed role such as, advisor, committee member, teacher, and mentor.

Princeton has provided an inexhaustible source of individuals who have shown unwavering support in many forms. Germán Rodríguez taught me graduate level statistics, provided statistical guidance on my research projects, and graciously agreed to be an examiner for my dissertation defense. Noreen Goldman helped me parse out epidemiology and demography. Joe Amon provided guidance in my graduate studies. The staff at OPR and EEB have provided me with support, friendships, and snacks. Thanks especially to Wayne Appleton, Elana Broch, Nancy Cannuli, Jenn Curatola, Mary Lou Delaney, Joann Donatiello, Lynne Johnson, Dawn Koffman, Joyce Lopuh, Judie Miller, Lil Morgan, Lolly O’Brien, Robin Pispecky, Diana Sacke, Val Smith, and Eleidis Sierra. I am thankful to my demography cohort Janeria Easley, Aaron Gottlieb, Mariana Campos Horta, Zitsi Mirakhur, Laura Nolan, and Melanie Wright Fox who made learning demography enjoyable, and whom I have gained valuable friendships. My OPR officemates, Theresa Andrasfay, Laura Blue, Lauren Gaydosh, Patrick Ishizuka, Sarah James, Vanessa Lehner, and Jess Yiu; and my Metcalf labmates, Ben Dalziel, Ayesha Mahmud, Joaquin Prada Jiménez de Cisneros, and Saki Takahashi
inspired me everyday to be a better researcher through their constant diligence and hard work. I was also fortunate to have Ayesha Mahmud as my partner in crime, as we made the transition from demography to disease modeling. “Senior” graduate OPR student gave me valuable feedback on my work, and have been a primary foundation from which I learned, especially Kristin Bietsch, Laura Blue, Ed Berchick, Diane Coffey, Audrey Dorélien, Dennis Feehan, Angelina Grigoryeva, Beth Sully, Takudzwa Sayi, Jonathan Tannen, Sal Thorkelson, and Megan Todd. I also truly appreciate the students of EEB, especially Christina Faust and Ruthie Birger, who accepted me as one of their own and made the transition to disease modeling feel within reach.

I am also grateful to those who taught and encouraged me even before I began the PhD. I am forever appreciative to my awesome wife Cathy who is the biggest supporter of my academic career. I am lucky to have loving parents, Joan and Dan Winter, who were my first, and most constant supporters. I am appreciative of my childhood friend, and fellow graduate school colleague, Jenny Vermilya, who was always a source of joy and laughter throughout our graduate student career. At Emory University, I am particularly grateful to my advisor Rob Stephenson who helped set my academic tract towards Princeton, and who has provided unwavering guidance to this day. Miriam Elfstrom, my friend and colleague at Emory University, shared my excitement in research and motivated me to pursue bigger and better things. At my undergraduate institution, the University of Georgia, I am thankful to Nan McMurry who patiently waded through my first research thesis with full encouragement. She graciously told me that my history thesis was actually a public health thesis, setting my career in motion.

Some additional acknowledgements, for each chapter, are in order. For chapter 1, I first want to thank India’s National Technical Advisory Group for reaching out to Jess Metcalf to explore rubella vaccine introduction, and Jess Metcalf for bringing me in on the project. I am grateful to Jess Metcalf, Bryan Grenfell, Justin Lessler, and Matt Ferrari for their helpful comments and discussion, and specifically Jess Metcalf and Justin Lessler for the use of their age-structured mathematical model. I also thank the Public Health Foundation of India, specifically Santanu Pramanik, Arpita Ghosh, Apoorva Sharan, Leena Inamdar, and Ramanan Laxminarayan who hosted my visit to Delhi and provided helpful comments for this research. Princeton University’s Center for Health
and Wellbeing, the Health Grand Challenges, funded my research trip to Delhi. I am grateful to
the participants, and particularly the discussant, Jeffrey Eaton, at the Population Association of
America annual conference who provided helpful feedback.

For chapter 2, I thank Jess Metcalf and Bryan Grenfell for their helpful comments and guidance,
and Jess Metcalf and Justin Lessler for the use of their age-structured model. I also thank Jess
Metcalf, Matt Ferrari, and Simon Cauchemez for motivating the need to this work, and additionally
Jess Metcalf and Amy Wesolowski for organizing the study and lab work. Many individuals at the
Institut Pasteur de Madagascar have worked tirelessly to analyze the serum samples, including
Miora Andrianjafimasy, Richter Razafindratsimandresy, Jean-Michel Heraud. I am also grateful to
Cara Brook, Malavika Rajeev, Rosamund Metcalf, and Peter Metcalf and for hosting an ushering
me around Madagascar, and aiding in my research.

For chapter 3, I thank Jess Metcalf and Bryan Grenfell for their helpful comments and guidance,
and Jess Metcalf and Justin Lessler for use of their age-structured model. I also thank Nicole Basta
for our multiple discussions about serology.

I am last thankful to all the individuals at the international, national, and local levels who work to
provide individuals with the life saving measles and rubella containing vaccines.
Contents

Abstract .................................................. iii
Acknowledgements ........................................ v
List of Tables ........................................... xi
List of Figures ........................................... xii

Introduction .............................................. 1

1 Rubella Vaccination in India: Identifying broad consequences of vaccine intro-
duction and key knowledge gaps ....................... 7

1.1 Abstract ............................................... 7
1.2 Introduction ......................................... 8
1.3 Methods ............................................... 11
  1.3.1 Simulation of Rubella Dynamics ................. 11
  1.3.2 Evaluation of Vaccination Strategies .......... 16
1.4 Results ................................................ 17
  1.4.1 Simulated Rubella Dynamics in India .......... 17
  1.4.2 Risk Factors for CRS Outbreaks ............... 25
1.5 Discussion ........................................... 33
Towards Measles Elimination: Exploring the Potential of Nested Serological Surveys within Measles and Rubella Surveillance Systems to Identify the Accumulation of Measles Susceptible Individuals in Madagascar

2.1 Abstract

2.2 Introduction

2.3 Methods

2.3.1 Data

2.3.2 Data Based on Non-Probability Sample

2.3.3 Direct Empirical Estimates of Proportion Seropositive by Age - Semiparametric Approach

2.3.4 Indirect Empirical Estimates of Proportion Seropositive by Age - Birth Cohort Projection of Immunity

2.3.5 Simulate Measles Dynamics - Age-Structured Model

2.3.6 Explore the Effect of Measles SIA Scenarios

2.4 Results

2.5 Discussion


3.1 Abstract

3.2 Introduction

3.3 Methods

3.3.1 Simulation of Three Endemic Rubella Populations

3.3.2 Serological Survey Design

3.3.3 Simulated Serological Surveys

3.3.4 Estimation of Age Specific Force of Infection - Three Analytic Approaches

3.3.5 Estimation of CRS Incidence Rate and $R_0$

3.4 Results

3.4.1 Analytic Method Bias

3.4.2 Survey Defined Age Group Bias
List of Tables

1.1 Summary of risk factors of CRS increase post-RCV introduction ........................................ 27
1.2 Summary of minimum required routine RCV vaccination coverage to prevent post-RCV introduction CRS increase, by $R_0$ .......................................................... 28
1.3 Estimated number of months until rubella outbreak post-RCV introduction, by $R_0$ and Indian state and region .......................................................... 31
1.4 Estimated number of months until CRS incidence rate $> 5$ cases per 100,000 live births post-RCV introduction, by $R_0$ and Indian state and region .................. 32

2.1 Percent reduction in the number of measles cases post-implementation of measles SIA scenarios .......................................................... 62

3.1 Types of parameter estimation biases from serological survey data ........................................ 70
3.2 Stepwise approach to exploring three types of parameter estimation biases from serological survey data .......................................................... 71

B.1 Model fits for the estimation of proportion measles seropositive in Madagascar, 2005-2015 ......................................................................................................................... 141
B.2 Model fits for the estimation of current proportion measles seropositive by age ..................... 143
B.3 Comparison between indirect and direct empirical estimates of proportion seropositive by age ......................................................................................................................... 144
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>India’s Population Age Structure, 2011</td>
<td>12</td>
</tr>
<tr>
<td>1.2</td>
<td>Crude birth rate and measles vaccination coverage, by Indian state and region</td>
<td>14</td>
</tr>
<tr>
<td>1.3</td>
<td>Estimated CRS incidence rate and incident cases, by Indian state and region</td>
<td>19</td>
</tr>
<tr>
<td>1.4</td>
<td>Estimated CRS incidence ratio over time for two Indian states</td>
<td>20</td>
</tr>
<tr>
<td>1.5</td>
<td>Estimated 30-year CRS incidence ratio, by Indian state and region, $R_0 = 5$</td>
<td>21</td>
</tr>
<tr>
<td>1.6</td>
<td>Estimated 30-year CRS incidence ratio taking into account private sector rubella vaccination, by Indian state and region, $R_0 = 5$</td>
<td>22</td>
</tr>
<tr>
<td>1.7</td>
<td>Estimated 30-year CRS incidence ratio <em>not</em> taking into account private sector rubella vaccination, by Indian state and region, $R_0$</td>
<td>23</td>
</tr>
<tr>
<td>1.8</td>
<td>Estimated CRS incidence rate over time in the presence and absence of RCV introduction taking into account predicted changes in the birth rate, rural Assam</td>
<td>24</td>
</tr>
<tr>
<td>1.9</td>
<td>Estimated CRS incidence rate over time in the presence and absence of RCV introduction keeping the birth rate constant over time, rural Assam</td>
<td>24</td>
</tr>
<tr>
<td>1.10</td>
<td>Estimated proportion seropositive by age and $R_0$ for one Indian region, 2015 and 2017</td>
<td>29</td>
</tr>
<tr>
<td>2.1</td>
<td>Madagascar’s population age structure and the sample age structure by five-year age groups</td>
<td>43</td>
</tr>
<tr>
<td>2.2</td>
<td>Sampling and vaccination ratios by region in Madagascar</td>
<td>44</td>
</tr>
<tr>
<td>2.3</td>
<td>Time series of measles and rubella cases, 2004-2015</td>
<td>54</td>
</tr>
<tr>
<td>2.4</td>
<td>Estimated population and under 5 years old proportion measles seropositive, 2005-2015</td>
<td>55</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>2.5 Direct and indirect estimates of proportion measles seropositive by age</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>2.6 Estimated population proportion measles seropositive post-implementation of measles SIA scenarios</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>2.7 Scatterplots of the estimated population proportion measles susceptible by measles $R_{eff}$ post-implementation of measles SIA scenarios, and by $R_0$</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>3.1 Simulated population age structure</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>3.2 Simulated populations age seroprevalence and age-specific force of infection</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>3.3 Epidemiological parameter estimates using perfect seroprevalence data by age in years, by analytic method, $R_0 = 5$</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>3.4 Epidemiological parameter estimates using perfect seroprevalence data by age in years, by analytic method, $R_0 = 3$</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>3.5 Epidemiological parameter estimates using perfect seroprevalence data by age in years, by analytic method, $R_0 = 7$</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>3.6 Two-age-group piecewise constant estimation of the age specific force of infection across three, two-age-groups</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>3.7 Epidemiological parameter estimates using perfect seroprevalence data by reproductive age in years, by analytic method, $R_0 = 5$</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>3.8 Epidemiological parameter estimates using perfect seroprevalence data by reproductive age in years, by analytic method, $R_0 = 3$</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>3.9 Epidemiological parameter estimates using perfect seroprevalence data by reproductive age in years, by analytic method, $R_0 = 7$</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>3.10 Estimation of age specific force of infection from proportion seropositive broken down into defined numerator and denominator</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>3.11 Epidemiological parameter estimates based on simulated serological data using fractional polynomials, $R_0 = 3$</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>3.12 Epidemiological parameter estimates based on simulated serological data using fractional polynomials, $R_0 = 5$</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>3.13 Epidemiological parameter estimates based on simulated serological data using fractional polynomials, $R_0 = 7$</td>
<td>103</td>
<td></td>
</tr>
</tbody>
</table>
3.14 Epidemiological parameter estimates based on simulated serological data using two-
age-group piecewise constants, $R_0 = 3$ ................................................. 104
3.15 Epidemiological parameter estimates based on simulated serological data using two-
age-group piecewise constants, $R_0 = 5$ ................................................. 105
3.16 Epidemiological parameter estimates based on simulated serological data using two-
age-group piecewise constants, $R_0 = 7$ ................................................. 106
3.17 Epidemiological parameter estimates based on simulated serological data using pe-
nalized regression splines, $R_0 = 3$ .................................................. 108
3.18 Epidemiological parameter estimates based on simulated serological data using pe-
nalized regression splines, $R_0 = 5$ .................................................. 109
3.19 Epidemiological parameter estimates based on simulated serological data using pe-
nalized regression splines, $R_0 = 7$ .................................................. 110
3.20 Global map of rubella-containing vaccine use ........................................ 120
3.21 Global map of crude birth rates .......................................................... 120

B.1 Predicted mean proportion of the population seropositive by age for the years (2005-
2015), based on model #4 ................................................................. 142
B.2 Time series of measles incidence: no SIA ........................................... 146
B.3 Time series of measles incidence: SIA target children ages 9 months old to 5 years
old and achieves 70% vaccination coverage ........................................ 147
B.4 Time series of measles incidence: SIA target children ages 9 months old to 5 years
old and achieves 80% vaccination coverage ........................................ 148
B.5 Time series of measles incidence: SIA target children ages 9 months old to 5 years
old and achieves 90% vaccination coverage ........................................ 149
B.6 Time series of measles incidence: SIA target children ages 9 months old to 5 years
old and achieves 95% vaccination coverage ........................................ 150
B.7 Time series of measles incidence: SIA target children ages 9 months old to 10 years
old and achieves 70% vaccination coverage ........................................ 151
B.8 Time series of measles incidence: SIA target children ages 9 months old to 10 years
old and achieves 80% vaccination coverage ........................................ 152
| B.9 | Time series of measles incidence: SIA target children ages 9 months old to 10 years and achieves 90% vaccination coverage | 153 |
| B.10 | Time series of measles incidence: SIA target children ages 9 months old to 10 years and achieves 95% vaccination coverage | 154 |
| B.11 | Time series of measles incidence: SIA target children ages 9 months old to 15 years and achieves 70% vaccination coverage | 155 |
| B.12 | Time series of measles incidence: SIA target children ages 9 months old to 15 years and achieves 80% vaccination coverage | 156 |
| B.13 | Time series of measles incidence: SIA target children ages 9 months old to 15 years and achieves 90% vaccination coverage | 157 |
| B.14 | Time series of measles incidence: SIA target children ages 9 months old to 15 years and achieves 95% vaccination coverage | 158 |
| B.15 | Time series of measles incidence: SIA target children ages 9 months old to 20 years and achieves 70% vaccination coverage | 159 |
| B.16 | Time series of measles incidence: SIA target children ages 9 months old to 20 years and achieves 80% vaccination coverage | 160 |
| B.17 | Time series of measles incidence: SIA target children ages 9 months old to 20 years and achieves 90% vaccination coverage | 161 |
| B.18 | Time series of measles incidence: SIA target children ages 9 months old to 20 years and achieves 95% vaccination coverage | 162 |
| B.19 | Time series of $R_{eff}$: no SIA | 164 |
| B.20 | Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 70% vaccination coverage | 165 |
| B.21 | Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 80% vaccination coverage | 166 |
| B.22 | Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 90% vaccination coverage | 167 |
| B.23 | Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 95% vaccination coverage | 168 |
B.24 Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 70% vaccination coverage .................................................. 169
B.25 Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 80% vaccination coverage .................................................. 170
B.26 Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 90% vaccination coverage .................................................. 171
B.27 Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 95% vaccination coverage .................................................. 172
B.28 Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 70% vaccination coverage .................................................. 173
B.29 Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 80% vaccination coverage .................................................. 174
B.30 Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 90% vaccination coverage .................................................. 175
B.31 Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 95% vaccination coverage .................................................. 176
B.32 Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 70% vaccination coverage .................................................. 177
B.33 Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 80% vaccination coverage .................................................. 178
B.34 Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 90% vaccination coverage .................................................. 179
B.35 Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 95% vaccination coverage .................................................. 180

C.1 Epidemiological parameter estimates for age groupings of ages 1-10 years using fractional polynomials, $R_0 = 3$ .................................................. 183
C.2 Epidemiological parameter estimates for age groupings of ages 1-10 years using fractional polynomials, $R_0 = 5$ .................................................. 184
<p>| C.3 | Epidemiological parameter estimates for age groupings of ages 1-10 years using fractional polynomials, $R_0 = 7$. | 185 |
| C.4 | Epidemiological parameter estimates for age groupings of ages 1-10 years using penalized regression splines, $R_0 = 3$. | 186 |
| C.5 | Epidemiological parameter estimates for age groupings of ages 1-10 years using penalized regression splines, $R_0 = 5$. | 187 |
| C.6 | Epidemiological parameter estimates for age groupings of ages 1-10 years using penalized regression splines, $R_0 = 7$. | 188 |
| C.7 | Epidemiological parameter estimates for age groupings of ages 1-10 years using two-age group piecewise constants, $R_0 = 3$. | 189 |
| C.8 | Epidemiological parameter estimates for age groupings of ages 1-10 years using two-age group piecewise constants, $R_0 = 5$. | 190 |
| C.9 | Epidemiological parameter estimates for age groupings of ages 1-10 years using two-age group piecewise constants, $R_0 = 7$. | 191 |
| C.10 | Epidemiological parameter estimates for age groupings of ages 10-20 years using fractional polynomials, $R_0 = 3$. | 193 |
| C.11 | Epidemiological parameter estimates for age groupings of ages 10-20 years using fractional polynomials, $R_0 = 5$. | 194 |
| C.12 | Epidemiological parameter estimates for age groupings of ages 10-20 years using fractional polynomials, $R_0 = 7$. | 195 |
| C.13 | Epidemiological parameter estimates for age groupings of ages 10-20 years using penalized regression splines, $R_0 = 3$. | 196 |
| C.14 | Epidemiological parameter estimates for age groupings of ages 10-20 years using penalized regression splines, $R_0 = 5$. | 197 |
| C.15 | Epidemiological parameter estimates for age groupings of ages 10-20 years using penalized regression splines, $R_0 = 7$. | 198 |
| C.16 | Epidemiological parameter estimates for age groupings of ages 10-20 years using two-age group piecewise constants, $R_0 = 3$. | 199 |
| C.17 | Epidemiological parameter estimates for age groupings of ages 10-20 years using two-age group piecewise constants, $R_0 = 5$. | 200 |</p>
<table>
<thead>
<tr>
<th>Page</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>Epidemiological parameter estimates for age groupings of ages 10-20 years using two-age group piecewise constants, $R_0 = 7$</td>
</tr>
<tr>
<td>203</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using fractional polynomials, $R_0 = 3$</td>
</tr>
<tr>
<td>204</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using fractional polynomials, $R_0 = 5$</td>
</tr>
<tr>
<td>205</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using fractional polynomials, $R_0 = 7$</td>
</tr>
<tr>
<td>206</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using penalized regression splines, $R_0 = 3$</td>
</tr>
<tr>
<td>207</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using penalized regression splines, $R_0 = 5$</td>
</tr>
<tr>
<td>208</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using penalized regression splines, $R_0 = 7$</td>
</tr>
<tr>
<td>209</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using two-age group piecewise constants, $R_0 = 3$</td>
</tr>
<tr>
<td>210</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using two-age group piecewise constants, $R_0 = 5$</td>
</tr>
<tr>
<td>211</td>
<td>Epidemiological parameter estimates for age groupings of ages 20-50 years using two-age group piecewise constants, $R_0 = 7$</td>
</tr>
<tr>
<td>213</td>
<td>Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using fractional polynomials, $R_0 = 3$</td>
</tr>
<tr>
<td>214</td>
<td>Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using fractional polynomials, $R_0 = 5$</td>
</tr>
<tr>
<td>215</td>
<td>Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using fractional polynomials, $R_0 = 7$</td>
</tr>
<tr>
<td>216</td>
<td>Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using penalized regression splines, $R_0 = 3$</td>
</tr>
<tr>
<td>217</td>
<td>Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using penalized regression splines, $R_0 = 5$</td>
</tr>
<tr>
<td>C.33 Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using penalized regression splines, $R_0 = 7$</td>
<td>218</td>
</tr>
<tr>
<td>C.34 Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using two-age group piecewise constants, $R_0 = 3$</td>
<td>219</td>
</tr>
<tr>
<td>C.35 Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using two-age group piecewise constants, $R_0 = 5$</td>
<td>220</td>
</tr>
<tr>
<td>C.36 Epidemiological parameter estimates for reproductive age groupings of ages 15-50 years using two-age group piecewise constants, $R_0 = 7$</td>
<td>221</td>
</tr>
<tr>
<td>C.37 Epidemiological parameter estimates for final four age groupings using fractional polynomials, $R_0 = 3$</td>
<td>223</td>
</tr>
<tr>
<td>C.38 Epidemiological parameter estimates for final four age groupings using fractional polynomials, $R_0 = 5$</td>
<td>224</td>
</tr>
<tr>
<td>C.39 Epidemiological parameter estimates for final four age groupings using fractional polynomials, $R_0 = 7$</td>
<td>225</td>
</tr>
<tr>
<td>C.40 Epidemiological parameter estimates for final four age groupings using two-age group piecewise constants, $R_0 = 3$</td>
<td>226</td>
</tr>
<tr>
<td>C.41 Epidemiological parameter estimates for final four age groupings using two-age group piecewise constants, $R_0 = 5$</td>
<td>227</td>
</tr>
<tr>
<td>C.42 Epidemiological parameter estimates for final four age groupings using two-age group piecewise constants, $R_0 = 7$</td>
<td>228</td>
</tr>
<tr>
<td>C.43 Epidemiological parameter estimates for final four age groupings using penalized regression splines, $R_0 = 3$</td>
<td>229</td>
</tr>
<tr>
<td>C.44 Epidemiological parameter estimates for final four age groupings using penalized regression splines, $R_0 = 5$</td>
<td>230</td>
</tr>
<tr>
<td>C.45 Epidemiological parameter estimates for final four age groupings using penalized regression splines, $R_0 = 7$</td>
<td>231</td>
</tr>
</tbody>
</table>
Introduction

Vaccination is one of the greatest public health interventions of the 20th century (CDC 1999). Vaccines have been effective at reducing the burden of infectious diseases. Noticeable vaccine triumphs include the eradication of smallpox in 1980 (World Health Organization and Global Commission for the Certification of Smallpox Eradication 1980), the eradication of rinderpest in 2010 (Office International des Epizooties 25 May 2011), and the elimination of endemic measles and rubella in the Americas region (CDC 2015, 2014). However, health policy decisions regarding vaccine programs are not always straightforward due to the complex transmission dynamics of infectious diseases. Even for pathogens with simple life-cycles, such as rubella and measles, which generate life long immunity from infection or vaccination, dynamics can be non-intuitive. Core considerations for predicting vaccination outcomes include vaccination coverage, population demography, and risk heterogeneities. For example, vaccination coverage at insufficient levels can increase the occurrence of late age infections, resulting in a different scope of potentially more damaging conditions. Examples include congenital rubella syndrome in infants born to mothers with rubella infection, or encephalitis from measles infections acquired in older children (Anderson and May 1983, Panagiotopoulos et al. 1999, Morice et al. 2003). Mechanistic epidemiological models provide essential tools for titrating these complexities. As such, mathematical models are increasingly relied upon for vaccination program design. For example, modeling has generated key insights into the outcomes of infectious disease control efforts (e.g., introducing the concept of herd immunity, or defining the changing epidemiology of infectious pathogen post vaccine introduction.
models are used to estimate burden of disease (e.g., estimating global CRS burden \cite{Vynnycky2016}, or Zika burden in the Americas region \cite{Perkins2016}), and to evaluate public health questions (e.g., progress towards measles mortality goals \cite{Simons2012}).

This dissertation uses mathematical modeling to address public health questions related to the current status of measles and rubella incidence and immunity, and the control of measles and rubella infections via vaccination.

Measles and rubella viruses share many similarities in terms of epidemiologically-relevant life histories; both pathogens are transmitted via respiratory droplets and direct contact, both are immunizing infections, both cause infections that result in symptoms of rash, fever, and cough, coryza, and/or conjunctivitis, and both have similar generation times \cite{WorldHealthOrganization2009,WorldHealthOrganization2011a}. Age-related changes in the force of infection are also well established for both infections and are generally attributed to changes in the pattern of contact over age \cite{Grenfell1985, Mossong2008, Anderson1983, Anderson1985}. Additionally, highly efficacious and safe vaccines, which mimic natural immunity resulting in life long immunization, are available for both measles and rubella \cite{Plotkin2013}. Measles vaccine, typically originating from the Edmonston strain, has a seroconversion rate of 86.9% among 8 to 9 month olds, and 99% among 11 to 12 month olds \cite{WorldHealthOrganization2009}. The most widely used rubella vaccine today is the human diploid cell rubella vaccine, RA 27/3, which has demonstrated a seroconversion rate of 95% or higher \cite{Beasley1969, WorldHealthOrganization2011a}. Both vaccines are typically administered to children. The first does of measles is typically administered to children between the ages of 9 and 12 months, and rubella is typically administered between 12 and 15 months. However, rubella-containing vaccine also has a history of being administered to women of reproductive age \cite{WorldHealthOrganization2009, WorldHealthOrganization2011a}. Given the shared pediatric vaccination schedule of the two vaccines and the fact that the vaccines are easily combined, there is a history of administering a combined vaccine within childhood immunization schedules (i.e., the
measles-rubella (MR) vaccine, the measles-mumps-rubella (MMR) vaccine, the measles-mumps-rubella-varicella (MMRV) vaccine). As a result, control measures for the two infections are often undertaken and discussed simultaneously (e.g., The Measles and Rubella Initiative (American Red Cross 2016)).

There are important differences between measles and rubella infections. The measles virus is more infectious and causes a more severe disease. Before the measles vaccine, more than 90% of individuals were infected by age 10 years old, the majority with symptoms (World Health Organization 2009); among children under five years old the case-fatality rate remains as high as 5 to 10% in some regions of the world (Wolfson et al. 2009). Rubella, on the other hand, is generally a mild infection in children, and between 20 and 50% of infections are asymptomatic (World Health Organization 2011a). However, rubella infection in pregnant women can cause detrimental outcomes such as spontaneous abortion, fetal death, and the birth of an infant with birth defects (i.e., congenital rubella syndrome (CRS)) (Greenberg et al. 1957; Enders et al. 1988; Miller et al. 1982). Despite great successes in reducing the burden of measles disease and CRS, twenty million individuals are infected with measles annually, an estimated 115,000 measles deaths occurred globally in 2014, and an estimated 105,000 CRS cases occur annually (Vynnycky et al. 2016). Given the continued burden of measles disease and CRS, there remain key public health questions regarding control of these two infections. As the epidemiological context changes, so will our reliance on mathematical models to anticipate and understand epidemiological shifts vital to the decision-making process for the measles and rubella control effort.

In addition to the broad dissertation theme which addresses public health questions related to the burden, control, and/or elimination of rubella and measles, three additional themes recur throughout. First, all chapters include a direct discussion and/or analysis of policy implications and recommendations for national vaccine control programs. Second, all chapters formally account for population demography. A plethora of literature has shown that demography, specifically the population age structure and rate of births for childhood infectious diseases, directly and
significantly affect transmission dynamics (John 1990b,a; Tuljapurkar and John, 1991; Manfredi and Williams 2004; Anderson and May 1991). For rubella, there is the additional core issue of the link between the age profile of fertility, average age of infection, and burden of disease. Third, all chapters include a focus on serological data, consisting of the need for and use of, and the strengths and limitations of this data source. Serological surveys are a powerful resource for understanding infectious disease dynamics. For immunizing infections like rubella and measles, this data source provides a window into the landscape of population immunity, i.e., the age distribution of individuals who are protected from infection. Serological surveys, in combination with mathematical models, can inform country specific knowledge gaps of disease burden and unknown transmission dynamics (Ximenes et al. 2014; Chakravarty et al. 1976), can be used to evaluate vaccination programs, determine if supplementary immunization activities are necessary and if so, the age group to be targeted, and even predict future outbreaks (Andrews et al. 2008; Babad et al. 1995).

**Prospectus for the dissertation**

In chapter 1, we evaluated the potential of introducing rubella-containing vaccine (RCV) into India’s public childhood vaccination schedule to reduce the burden of CRS. We originally evaluated this question because we were charged with the task by India’s National Technical Advisory Group on Immunizations. We used a deterministic age-structured model that accounts for state-specific demography and vaccination coverage levels to simulate rubella dynamics in the absence and presence of the introduction of RCV. We then compared simulated output and found that introduction of RCV via an initial catch-up campaign of 60% coverage among children 9 months to 14 years old followed by routine immunization campaign at coverage levels equal to estimated measles coverage will result in a 30-year cumulative decrease in CRS incidence in all states and regions if the rubella basic reproductive number, a summary measure of transmission, is nine or lower. We found that RCV introduction will be most successful at reducing the cumulative burden of CRS in states with falling birth rates. However, our findings also indicated that the effect of introducing RCV
on the transient annual incidence of CRS is highly sensitive to the the basic reproductive number of rubella. Therefore, we identified risk factors that can be used to highlight regions at risk of transient increases in CRS burden post-RCV introduction. We conclude that the results provide a generally optimistic view of the introduction of the rubella vaccine in India such that cumulative CRS incidence is expected to decrease; however, an investment in rubella surveillance will be vital to monitor the success and possible failures of the vaccination programs. A previous version of chapter 1 was presented at the 2016 Population Association of America Annual Meeting on April 2, 2016 in Washington, DC.

In chapter 2, we explored the use of a nested serological survey within the measles and rubella surveillance system in Madagascar in order to estimate population immunity in the context of post-vaccine measles dynamics. We then evaluated the potential of measles Supplementary Immunization Activities (SIA) to reduce the risk of a potentially looming measles outbreak in Madagascar. We analyzed 4394 serum samples that were collected between November 2004 and December 2015 from the measles and rubella surveillance system in Madagascar. We used semi-parametric models with penalized regression smoothers to directly estimate the current proportion of seropositive individuals by age. To evaluate the strength of this inference, we compared direct estimates of seroprevalence by age to indirect estimates based on an age cohort projection of immunity based on methods by Takahashi et al. (2015). We estimated that only 84.9% of the total Malagasy population is immune to measles (95% CI: 75.3%, 89.8%), indicating that Madagascar is at risk of major outbreak. The indirect estimate fell just inside the 95% confidence bounds of this point estimate; however, the two age profiles of seroprevalence differed markedly over age. We additionally predicted the effect of multiple SIA scenarios on the risk of a major measles outbreak. By comparing estimates based on a nested serological survey within the measles and rubella surveillance system to indirect estimates, this analysis took the first step towards recommending this low-cost data source for general use. Portions of this chapter were presented at the Institute for Disease Modeling Symposium on April 20, 2016 in Bellevue, WA.
In chapter 3, we analyzed the strengths and limitations of rubella IgG serological data to characterize rubella epidemiological parameters and CRS incidence, which determine the effect of rubella vaccine introduction. We used an established age-structured rubella model to simulate three populations representing a range of rubella transmission levels, and extracted variables of interest: the age immunity profile, the age specific force of infection, rubella’s basic reproductive number, and CRS incidence rate. We then systematically explored biases from i) analytic methods, ii) chosen survey age groups, and iii) sampling bias, by comparing ‘true’ parameter values (extracted from simulated populations) to parameter estimates. We laid out in detail nuanced relationships between these characteristics and bias that can be used by public health officials to understand past serological survey estimates. We additionally recommended rubella serological survey designs to be used to evaluate current CRS burden and the basic reproductive number of rubella to aid in decision for introduction of rubella vaccine.

Finally, we conclude with a brief summary and a discussion of a few of the many directions that future research can take.
Rubella Vaccination in India: Identifying broad consequences of vaccine introduction and key knowledge gaps

1.1 Abstract

Introduction: Rubella is predominantly a mild childhood infectious disease. However, rubella virus infection during early pregnancy may cause fetal death, spontaneous abortion, or the birth of an infant with congenital rubella syndrome (CRS). In 2014, India announced their plan to introduce rubella-containing vaccine (RCV) into India’s national measles immunization program. The success of India’s rubella vaccination program will depend on maintaining vaccination coverage above a critical threshold. Empirical analyses have demonstrated that low levels of RCV coverage can result in an increase in CRS incidence in the short-term, by increasing the age of infection without sufficiently reducing incidence of rubella cases. Methods: We used a deterministic age-structured model that accounts for Indian state and rural and urban region-specific demography and vaccination coverage levels to simulate rubella dynamics in absence and presence of introducing RCV within the public sector. We then compared CRS incidence over time between the different simulations. Results: We found that the introduction of the RCV via an initial catch-up campaign with 60% coverage among children 9 months to 14 years old, and routine immunization campaign at coverage levels equal to current measles coverage will result in a 30-year cumulative decrease in Congenital Rubella Syndrome incidence in all states if the rubella basic reproductive number, $R_0$, is nine or lower. RCV introduction will be most effective in reducing the cumulative burden of CRS in states with falling birth rates. However, our results indicate that the effect of introducing RCV on the transient annual incidence of CRS is highly sensitive to the $R_0$ for rubella. As a
result, we identify risk factors that can be used to highlight regions at risk of transient increases in CRS burden post-RCV introduction. **Conclusions:** These results provide a generally optimistic view of the introduction of the rubella vaccine in India; the rubella vaccine will likely result in cumulative decreases in CRS. However, an investment in rubella surveillance will be vital to monitor the success and possible failures of vaccination programs in order to prevent rubella outbreaks that can result in annual increases in CRS. To this point, population-based serological surveys of rubella, and improved CRS surveillance and fever/rash measles and rubella surveillance can be used to estimate population immunity, evaluate the effectiveness of previous vaccine control efforts at reducing population susceptibility, and evaluate the need for Supplementary Immunization Activities (SIAs). Additionally, India must maintain a readiness and flexibility to administer SIAs, specifically in identified states with low vaccination coverage.

### 1.2 Introduction

Rubella typically presents as a mild febrile rash illness in children. However, rubella infection in pregnant women can cause detrimental outcomes such as spontaneous abortion, fetal death, and the birth of an infant with birth defects (i.e., congenital rubella syndrome (CRS)) (Greenberg et al., 1957; Enders et al., 1988; Miller et al., 1982). Rubella-containing vaccine (RCV) is a safe and effective vaccine (Reef and Plotkin, 2013) that with high uptake can be used to successfully eliminate endemic rubella and CRS cases as has occurred in the Americas region (Figueroa et al., 2014). As of 2015, RCV had been introduced into the national immunization schedules in 140 of 194 WHO member states (CDC, 2015). The majority of CRS cases occur in the World Health Organization (WHO) regions of South East Asia and Africa, where uptake of RCV is lowest (Vynnycky et al., 2016). Vynnycky et al. (2016) estimated that 49,000 of the 105,000 2010 incident cases of CRS occur in WHO’s South-East Asia Region, and the majority of the cases (40,000) were estimated to occur among Indian infants.

In 2014, the government of India announced that they plan to introduce RCV into India’s childhood Universal Immunization Programme (UIP) (British Broadcasting Company, 2014; Hindustan Times, 2014). Uptake of RCV in India has the potential to substantially reduce the global burden of CRS. However, research has shown that RCV coverage below a critical threshold can actually result in an increase in CRS incidence beyond rubella endemic CRS levels. Inadequate childhood
vaccination coverage can increase the age of infection without sufficiently reducing incidence of rubella cases such that females, who would likely have contracted rubella in childhood years prior to vaccine introduction, now age into their reproductive years still susceptible to rubella and at risk of giving birth to a child with CRS [Anderson and May, 1983; Knox, 1980; Panagiotopoulos et al., 1999; Morice et al., 2003; Metcalf et al., 2012b]. The undesired outcome of increasing CRS incidence after RCV introduction can be evaluated by: i) comparing cumulative 30-year CRS estimates in the vaccine and non-vaccine prediction scenarios, or ii) comparing transient or annual CRS estimates in the vaccine and non-vaccine prediction scenarios.

Ultimately, the success of India’s rubella vaccination program will depend on reaching a critical threshold of vaccination coverage to prevent cumulative increases in CRS burden, and prevent transient increases in CRS incidence. An increase in the cumulative burden of CRS as a result of a RCV introduction program would be considered a major public health programatic failure. Prevention of this outcome is necessarily considered priority number one, however there is no evidence of a sustained CRS increase resulting from a real world vaccination to date (Anderson and May, 1983; Knox, 1980). A secondary RCV introduction program goal is to prevent transient increases in CRS due to late age outbreaks of rubella. Empirical data from Costa Rica and Greece have demonstrated these transient or short-term increases in CRS after introduction of RCV into each country’s respective childhood immunization schedules [Panagiotopoulos et al., 1999; Morice et al., 2003; Metcalf et al., 2012b]. Short-term CRS increases are not an inevitable outcome of RCV introduction; program evaluation and surveillance and supplementary immunization activities (SIAs) can be used to avoid transient increases in CRS burden.

The level of vaccination coverage required to avoid an increase in CRS cases depends predominantly on three population level characteristics: population birth rate, transmissibility of disease, and whether vaccination has interrupted endemic transmission. Previous mathematical models have indicated that for childhood immunization programs, at least 80% immunization coverage of one dose of RCV is sufficient to avoid a post-vaccination increase in CRS incidence (Anderson and May,
The WHO stated in their 2011 rubella position paper “to avoid the potential of an increased risk of CRS, countries should achieve and maintain immunization coverage of 80% or greater with at least 1 dose of an RCV delivered through routine services or regular SIAs, or both” (World Health Organization, 2011a). Since 2011, the 80% vaccination coverage recommendation has been confirmed as a conservative estimate across a range of demographic and epidemiological contexts (Lessler and Metcalf, 2013; Metcalf et al., 2012a). Assuming current measles vaccine coverage rates in India, RCV coverage will likely range between 60% and 90% across states (Office of the Registrar General & Census Commissioner, 2012a). Based on Metcalf et al. (2013), assuming assortative age mixing patterns, some Indian states with low coverage may be at risk of increasing CRS incidence over the long-term if RCV is introduced into India’s national measles childhood immunization program. No previous research has assessed the critical vaccination coverage to prevent any transient CRS incidence increases.

The objective of this study was to explore the effect of introducing RCV into India’s public sector immunization program on CRS incidence in both the short and long-term. Using a deterministic age-structured mathematical model, we simulated rubella dynamics in the Indian population in the presence and absence of the introduction of RCV into the public sector, and assessed the effect of different vaccination strategies on CRS incidence. This analysis builds on previous work exploring RCV introduction by: 1) assessing the effect on both long-term cumulative CRS burden and short-term transient annual CRS burden, 2) allowing birth rates to change over time, 3) assessing the potential role of private sector rubella vaccination, and 4) leveraging within country vaccine and demographic heterogeneity. We discuss key assumptions underlying the analyses, conduct sensitivity analysis of unknown parameters, and identify risk factors that can be used to anticipate potential transient increases in CRS incidence post vaccine introduction.
1.3 Methods

1.3.1 Simulation of Rubella Dynamics

We used an established age-structured mathematical model to simulate rubella transmission dynamics among the Indian population. The key feature of the model was a matrix that at every time-step defines transition from every possible epidemiological stage (e.g., maternally immune, susceptible, infected, recovered, and vaccinated) and age combination to every other possible epidemiological stage and age combination. The discrete time-step was set to two weeks based on the generation time of rubella, thus there were 24 opportunities to move in and out of age and epidemiological stages per year. At every biweekly time period, the simulation outputs the number of individuals in each age and epidemiological stage. The model methods are presented in greater detail in Metcalf et al. (2012b) and Metcalf et al. (2012a), and are therefore not further discussed here.

Necessary data inputs required by the age-structured model in order to simulate the time-course of incidence of rubella in the population included: known aspects about the epidemiology of rubella, India’s demographics, and predicted levels of rubella vaccination coverage. The basic reproductive number, $R_0$, is defined as the average number of people a typical infected individual will infect in a fully susceptible population. Estimates of $R_0$ for rubella based on empirical data have ranged from 3 in Europe to 12 in Ethiopia (Anderson and May, 1991; Cutts et al., 2000; Edmunds et al., 2000; Fine, 1993). Unfortunately, we have found no published study that estimates the $R_0$ for rubella in India. Given this serious lack of data and knowledge, we assumed an $R_0$ of 5 based on a previous in-depth analysis of 40 African countries (Lessler and Metcalf, 2013), and additionally ran a sensitivity analysis for larger $R_0$ values including 7, 9, and 11, which result in more conservative effects of vaccine. We also assumed a pattern of contact over age proportional to what was measured in Great Britain based on a diary study in Europe (Mossong et al., 2008), and assumed that transmission fluctuates seasonally, as has been reported for rubella in an array of settings (Keeling et al., 2001; Metcalf et al., 2011b; Wesolowski et al., 2015; Metcalf et al., 2011a).
We simulated rubella transmission dynamics from 1991 to 2046 in India’s population. The simulations ran for 25 years (1991 to 2016), at which time we allowed a range of vaccination strategies to be introduced and then ran the simulation for another 30 years (2016-2046). We simulated rubella transmission dynamics for 20 years prior to any experiments in order to move beyond the transient non-seasonal outbreaks while maintaining appropriate population structure (Figure 1.1). Simulations were run independently for all Indian states by their rural and urban regions (with the exceptions of Goa and Sikkim) for a total of 52 regions. The two excluded states introduced RCV into their public sector immunization program prior to 2016, and therefore, RCV coverage is unknown.

Figure 1.1: India 2011 age structure: comparing simulated population to census population, males and females.

Our primary objective was to explore the impact of different vaccination scenarios on the burden of CRS. In scenario one, the ‘no vaccine’ scenario, we simulated rubella disease dynamics through time assuming no vaccination in the public or private sector. In scenario two, the ‘no public sector’ vaccine scenario, we simulated rubella disease dynamics through time allowing low level private
sector rubella vaccination. In scenario three, or the ‘routine’ vaccine scenario, we introduced RCV into the routine childhood measles national immunization program in 2016 at levels defined for each state by region. In scenario four, or the ‘routine + catch-up’ vaccine scenario, we introduced RCV as a MR catch-up campaign for ages 9 months to 14 years old in 2016, assuming a conservative estimate of 60% coverage (Sharan, 2014), and added RCV into the routine childhood measles immunization program. Each scenario was simulated independently for all 52 regions.

Population survival rates per age class were estimated for all 52 regions by matching the region’s estimated life expectancy from 2010 to a United Nations model ‘Coale-Demeny East’ life table (Office of the Registrar General & Census Commissioner, 2012b; UN Population Division, 2010). Life Expectancy for rural and urban Uttarakhand and the north eastern states with the exception of Assam (i.e., Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Tripura) assumed the life expectancy of rural and urban India, respectively. We extracted population totals and age structures for all 52 regions from the 1991 Indian Census (Office of the Registrar General & Census Commissioner, 1991), and rescaled the population in 2011 to match the 2011 Indian Census (Office of the Registrar General & Census Commissioner, 2011). Smooth splines were used to estimate survival rates and the population structure by ages in month intervals, rather than by year intervals.

Crude birth rates (CBR) from 1971 through 2013 were extracted for all 52 regions from the Sample Registration System (Office of Registrar General & Census Commissioner, 1998, 1999-2014). The CBR for Chhattisgarh, Jharkhand, and Uttarakhand prior to 1999 were input from Madhya Pradesh, Bihar, and Uttar Pradesh CBR, respectively. India is in phase II of the fertility transition which is falling fertility (United Nations, 2015; United Nations and Department of Economic and Social Affairs and Population Division, 2015). We fit a negative sigmoid function to the CBR data over time. If by 2013, the region’s births had not yet dropped below replacement level total fertility rate (TFR) of 2.1, we estimated the lower asymptote as the minimum CBR equal to replacement level TFR of 2.1 times the average number of women per age year between 15 and 40 years old divided by the population total. The average number of women and the population total was
based on the 2011 census for each region (Office of the Registrar General & Census Commissioner, 2011). If the region’s TFR fell below replacement level of 2.1, then the lower asymptote CBR was set to the recent CBR minus one. Figure 1.2 displays the estimated 2011 crude birth rates for each region. Declining birth rates in the absence of vaccine are expected to increase the CRS burden (Metcalf et al., 2012a), so including this decline was important to evaluating the impact of vaccination on the burden of CRS in a realistic India population.

Figure 1.2: State level covariates. a) Urban 2011 crude birth rates per 1,000 people. b) Rural 2011 crude birth rates per 1,000 people. c) Urban 2011 measles vaccination coverage rates (as a proportion). d) Rural 2011 measles vaccination coverage rates (as a proportion).

Age-specific fertility rate (ASFR) estimates were extracted from the 2012 Sample Registration System (Office of Registrar General & Census Commissioner, 2012). Due to missing data, the fol-
lowing small states took on India’s average rural ASFR and urban ASFR for their rural and urban counterparts: Uttarakhand and the north eastern states with exception of Assam, i.e., Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Tripura. In the absence of information on the details of how fertility decline might affect the age pattern of fertility, the relative age pattern was kept constant over the course of the simulation, but the magnitude was adjusted based on the birth rate for each time step to estimate CRS cases, details are provided below.

As it is likely that rubella will be distributed as part of a bivalent measles-rubella (MR) vaccine in India (Sharan, 2014), we used estimates of measles-containing vaccine dose 1 (MCV1) routine vaccination coverage as a basis for levels of rubella routine immunization coverage implemented in the model. Specifically, rubella routine immunization coverage rates in the two main vaccine scenarios (‘routine’ and ‘routine + catch-up’) were set to match the 2011 MCV1 routine coverage rates estimated by Pramanik et al. (2015). Figure 1.2 shows the estimated MCV1 coverage rates by Indian state and region. As no state level measles vaccination coverage rates have been published since 2007, these estimates from Pramanik et al. (2015) rely on inference from data taken across multiple years and surveys. In the absence of further information, we assumed these rates remained constant over time.

Vynnycky et al. (2003) found that the magnitude of private sector rubella vaccination can affect the outcome of public sector vaccine introduction. We therefore conducted a sensitivity analysis allowing for low level private sector rubella vaccination via scenario two as discussed above. In addition to comparing the two vaccine scenarios (‘routine’ and ‘routine + catch-up’) to ‘no vaccine’ scenario, we also compare the two vaccine scenarios (‘routine’ and ‘routine + catch-up’) to the ‘no public sector’ vaccine scenario. There is minimal published data on private sector rubella vaccine; estimates of private sector RCV vaccine coverage have been estimated as high as 40% in Delhi (Dewan and Gupta, 2012). We explored a range of potential private sector RCV coverage levels among children ages 9 to 24 months old of 5%, 15%, 25%, and 35%. The sensitivity analysis assumed these private sector RCV coverage rates affected all birth cohorts beginning in 2011. For
simplicity, and in the absence of detailed information on this, we assumed that populations with
access to private sector vaccine are well mixed with those that are not, a reasonable assumption
for an infection as transmissible as rubella.

1.3.2 Evaluation of Vaccination Strategies

We evaluated the effects of RCV introduction into India’s national measles immunization program
by calculating the number of CRS cases and CRS incidence rates for each scenario. We also
estimated CRS incidence ratios in order to compare the ‘no vaccine’ scenario and the ‘no public
sector’ vaccine scenarios to the ‘routine’ and ‘routine + catch-up’ vaccine scenarios. The simulations
ran for 25 years (1991 to 2016), at which time the vaccine was introduced and then the simulation
ran for another 30 years (2016-2046). The annual number of CRS cases and the CRS incidence per
100,00 live births for each age was estimated for each Indian state by rural and urban regions for
each time period between 2011 and 2046. We assumed that women who are susceptible have the
same age-specific fertility rate as non-susceptible women, where the incidence of becoming infected
during pregnancy was equal to susceptible non-pregnant women becoming infected. The number
of CRS cases for each time $t$ is:

$$\text{CRScases}(t) = \sum_{a=15}^{a=49} S(a, t) \times i(a, t) \times f(a) \times \nu(t) \times 0.65$$  \hspace{1cm} (1.1)

where $S(a, t)$ is the number of susceptible individuals at each age $a$ and time $t$, $i(a, t)$ is the
probability of becoming infected with rubella over a 16 week period for each time $t$ and age $a$,
and $f(a)$ is the age-specific fertility rate per 1 woman. The $\nu(t)$ term is a scalar; it is the ratio
of the total number of births estimated from the crude birth rate at time $t$ divided by the total
number of births estimated from the age-specific fertility rate at time $t$. By maintaining a constant
age-specific fertility rate, we assumed no change in the proportion of births by women’s age, while
allowing the total number of births to decrease over time. We also assumed that the probability of
CRS following rubella infection during the first 16 weeks of pregnancy was 0.65 \cite{Vynnycky et al.}. 

16
The following is the CRS incidence per 100,000 live births for each time period $t$.

$$CRS_{\text{incidence}}(t) = \frac{CRS_{\text{cases}}(t)}{\text{births}(t)} \times 100,000 \quad (1.2)$$

Long-term CRS incidence rate was estimated by taking the average CRS incidence rate between 2016 and 2046. Short-term CRS incidence rates were assessed by estimating the CRS incidence rate each year between 2016 and 2046. The CRS incidence ratio is a measure of the effect of vaccination introduction on CRS. It is defined as the incidence of CRS in the reference ‘no vaccine’ scenario divided by the incidence of CRS in a vaccine scenario (‘routine’ and ‘routine + catch-up’). With the exception of the private sector sensitivity analysis, the reference simulation is always the ‘no vaccine’ scenario. A CRS incidence ratio of greater than one means that the risk of CRS actually increased as a result of vaccination introduction. The CRS incidence ratio was calculated each year for all 52 regions.

$$CRS_{\text{incidence\_ratio}}(t) = \frac{CRS_{\text{incidence\_vaccine}}(t)}{CRS_{\text{incidence\_novaccine}}(t)} \quad (1.3)$$

The long-term CRS incidence ratios were estimated using the 30-year average CRS incidence rates in each scenario. The short-term CRS incidence ratios were estimated by comparing each year’s CRS incidence rates across vaccine scenarios. All statistical analyses were done using R 3.2.2 (R Core Team, 2016).

### 1.4 Results

#### 1.4.1 Simulated Rubella Dynamics in India

The simulation results suggest that approximately 15,000 children were born with CRS in India in 2015, an incidence rate of 56 CRS cases per 100,000 live births. This estimated number of CRS cases falls at the lower end of the 2010 range estimated by Vynnycky et al. (2016) of an average of 40,000 CRS cases (ranging from 0 to 86,000). Figure 1.3 displays the CRS incidence per 100,000 live births and the number of CRS cases for each state, by rural and urban regions. Rural Andhra Pradesh, and West Bengal (95 and 88 per 100,000 live births respectively) have the
highest CRS incidence rates; rural Uttar Pradesh and rural Bihar (34 and 37 cases per 100,000 live births respectively) have the lowest CRS incidence rates. A comparison of Figures 1.2a and 1.2b to Figures 1.3a and 1.3b indicates that the states with the highest birth rates have the lowest CRS incidence rates in 2015. Higher birth rates means that the number of susceptible individuals entering the population is higher, therefore there is more transmission of rubella in the population, and a shorter wait before an individual acquires infection, thus decreasing the average age of infection and the risk of CRS. Despite low CRS incidence, the large population of Uttar Pradesh carries the largest burden of total number of CRS cases in its rural population with over 1,500 cases of CRS in 2015, followed by rural Bihar with 1,006 cases.

As described above, our primary objective was to explore the impact of different vaccination scenarios on CRS incidence. Accordingly, we further simulated rubella dynamics for rural and urban regions in 52 Indian states under four different vaccine scenarios: 1) no RCV in the public or private sector, 2) RCV only in the private sector, 3) RCV as a routine immunization only, and 4) RCV as routine immunization and catch-up campaign. The vaccine was introduced in 2016 in our simulations. Figure 1.4a displays the simulated results for urban Maharashtra, which represents the typical pattern we see across most regions assuming an $R_0$ of five. Figure 1.4a shows that, compared to the ‘no vaccine’ simulation, in the presence of vaccination, the number of CRS cases each year drops sharply. Approximately 10 years post-RCV introduction, the CRS incidence falls off completely in the ‘routine’ scenario (red dots), and falls off almost immediately in the ‘routine + catch-up’ scenario (green dots). On average, the CRS incidence ratios for the ‘routine’ vaccine scenario compared to the ‘no vaccine’ scenario and for the ‘routine + catch-up’ vaccine scenario compared to the ‘no vaccine’ scenario were less than one in urban Maharashtra (0.1217 and 0.0003 respectively). In fact, the average CRS incidence ratio between 2016 and 2046 for both the ‘routine’ and ‘routine + catch-up’ vaccine scenarios compared to ‘no vaccine’ scenario was less than one for all 52 regions (Figure 1.5). Therefore, our results indicate that if $R_0$ for rubella in India is five, the introduction of RCV into the national public sector (via ‘routine’ or ‘routine + catch-up’ vaccine scenarios) will result in a long-term decrease in CRS incidence over 30-years in
Figure 1.3: Results of simulated rubella dynamics assuming an $R_0$ of five and no private sector vaccination. a) Urban 2015 estimated CRS incidence per 100,000 live births by state. b) Rural 2015 estimated CRS incidence per 100,000 live births by state. c) Urban 2015 estimated number of CRS cases by state. d) Rural 2015 estimated number of CRS cases by state.

all simulated states in India.

The 30-year average long-term CRS incidence ratio ignores short-term rubella transmission dynamics, which may result in short periods of increased numbers of CRS cases relative to those experienced in the absence of vaccination. Assuming a $R_0$ of five, there are a few states (Arunachal Pradesh, Nagaland, Uttar Pradesh) whose transient dynamics of rubella warrant further attention. Figure 1.4b shows the CRS incidence ratio over time for rural Uttar Pradesh, which represents the pattern generally seen in these three states. The incidence of CRS falls off immediately in
Figure 1.4: Results of simulated rubella dynamics assuming an $R_0$ of five and no private sector vaccination. a) The CRS incidence ratio 2011 to 2046 in rural Maharashtra, a pattern seen in the majority of states. b) The CRS incidence ratio 2011 to 2046 in rural Uttar Pradesh, a pattern seen in Arunachal Pradesh and Nagaland as well. The reference simulation is the ‘no vaccine’ scenario represented by the blue horizontal line at CRS incidence ratio=1, the CRS incidence ratio of the ‘routine’ vaccine scenario compared to the ‘no vaccine’ reference scenario is represented by red dots, and the CRS incidence ratio of the ‘routine + catch-up’ vaccine scenario compared to the ‘no vaccine’ reference scenario is represented by green dots.

the ‘routine + catch-up’ vaccine scenario, and ten years later increases substantially but remains below one for the remainder of the simulation. Within the ‘routine’ vaccination only scenario, the incidence of CRS decreases for the first ten years post RCV introduction, and then begins to increase again but never reaches baseline levels. Smaller decreases in CRS incidence in these states are due to estimated low vaccination coverage rates ($<70\%$) that allow continuing circulation of rubella, which in turn may result in outbreaks associated with infection of susceptible women too old to have been reached by the vaccination program, or girls missed by the vaccination program who have had time to age into their childbearing years. Despite rubella outbreaks post vaccine introduction that may be associated with changes in the numbers of CRS cases, on average over 30 years, our model results indicate that these states will also see an overall reduction in rates of CRS given vaccine introduction. Furthermore, our results show the CRS incidence ratio for these states remains below one over time, suggesting that current vaccination coverage rates are sufficient to avoid even transient increases in CRS incidence.
Figure 1.5: Results of simulated rubella dynamics assuming an $R_0$ of five and no private sector vaccination: 30-year average CRS incidence ratio between 2016 and 2046 across all states by region and vaccination scenario. a) Urban CRS incidence ratio of ‘routine’ vaccine scenario compared to ‘no vaccine’ scenario. b) Urban CRS incidence ratio of ‘routine + catch-up’ vaccine scenario compared to ‘no vaccine’ scenario. c) Rural CRS incidence ratio of ‘routine’ vaccine scenario compared to ‘no vaccine’ scenario. d) Rural CRS incidence ratio of ‘routine + catch-up’ vaccine scenario compared to ‘no vaccine’ scenario. Introduction of RCV into the national public sector (via ‘routine’ or ‘routine + catch-up’ vaccine scenarios) will result on average over 30 years in a decrease in CRS incidence in every simulated state in India.

While private sector vaccination alone may slightly increase the burden of CRS in the absence of public sector vaccine introduction, we find that the presence or absence of private sector vaccination does not affect our inference as to the effect of RCV introduction into the public sector. The results of a sensitivity analysis for the ‘no public sector’ vaccine revealed that when
we allow for low-level private sector RCV coverage among children ages 9 to 24 months ranging from 5-35%, the CRS incidence ratio in both vaccine scenarios (‘routine’ and ‘routine + catch-up’) compared to the ‘no public sector’ vaccine scenario decreases slightly (Figure 1.6). Therefore, by assuming zero private sector rubella vaccination in the simulations above, we are conservatively underestimating the decrease in CRS incidence by introduction of RCV into the public sector. We find that we can therefore afford to neglect private sector vaccination, a useful result given the paucity of information on private sector vaccination coverage.

Figure 1.6: Results of simulated rubella dynamics assuming an $R_0$ of five: average CRS incidence ratio between 2016 and 2046 across all states by region. The top row is urban CRS incidence ratio of ‘routine + catch-up’ vaccine scenario compared to ‘no public sector’ vaccine scenario with 5% private sector vaccine coverage, 15% private sector vaccine coverage, 25% private sector vaccine coverage, and 35% private sector vaccine coverage from left to right. The bottom row is rural CRS incidence ratio of ‘routine + catch-up’ vaccine scenario compared to ‘no public sector’ vaccine scenario with 5% private sector vaccine coverage, 15% private sector vaccine coverage, 25% private sector vaccine coverage, and 35% private sector vaccine coverage left to right. Allowing for low level private sector RCV coverage among children ages 9 to 24 months, the CRS incidence ratio in the ‘routine + catch-up’ vaccine scenario compared to the ‘no public sector’ vaccine scenario slightly decreases, although there are no noticeable differences in the figure. Therefore, by not taking into account private sector vaccine, the results slightly underestimate the decrease in CRS incidence by introduction of RCV into the public sector.

The results discussed above were based on simulations that assumed a rubella $R_0$ of five. Figure 1.7 displays the results of a sensitivity analysis for an increased range of values of $R_0$. For simplicity, we present average CRS risk ratio over 30 years comparing the ‘routine + catch-up’ vaccine scenario to the ‘no public sector’ vaccine scenarios, as this presents the most likely scenario of RCV introduction in India (Sharan, 2014). Figure 1.7 shows the CRS incidence ratio is sensitive to the
value of $R_0$. As $R_0$ increases, the CRS incidence ratio increases as well, and our confidence in the successful introduction of RCV decreases given current vaccine levels. It is important to note that in the absence of vaccination, higher values of $R_0$ are expected to result in a lower incidence of CRS (Metcalf et al., 2012a). Therefore, smaller numbers of CRS cases in simulations with higher $R_0$ values can result in larger CRS incidence ratios, even though the size of the outbreak may be very small (Figure 1.8).

The transient dynamics of CRS over time are also highly sensitive to the assumed value of $R_0$. A higher $R_0$ results in earlier outbreaks of rubella and associated CRS cases post vaccine introduction (see Winter (2016) for link to online figure). But it is worth noting that in transient outbreaks as well, higher values of $R_0$ are expected to result in a lower incidence of CRS (Metcalf et al., 2012a).

The model allowed birth rates to change over time rather than remaining constant as previously assumed when evaluating RCV introduction (Metcalf et al., 2012a). We assumed that for regions
above replacement fertility levels, birth rates would fall to replacement level. Declining birth rates results in an increase in the number of CRS cases given an increase in the average age of infection (Figure 1.8 vs. 1.9), the vaccine can then have a larger impact on reducing the cumulative burden of CRS than if birth rates were declining. Therefore, RCV introduction will be most effective in reducing the cumulative burden of CRS in regions with falling birth rates.

Figure 1.8: Results of simulated rubella dynamics assuming no private sector vaccination. The figure shows the CRS incidence per 100,000 live births for rural Assam from 2016 to 2046 by $R_0$. The solid lines show the CRS incidence in the ‘routine + catch-up’ vaccine scenario for different assumed initial values of $R_0$. The dotted lines shows the CRS incidence over time in the ‘no vaccine’ scenario for different assumed initial values of $R_0$. The incidence of CRS is higher for lower values of $R_0$ in the ‘no vaccine’ scenario, which is a consistent finding across states. As a result, smaller CRS incidence rates in the high $R_0$ scenario can result in larger CRS incidence ratios, even though CRS incidence may be smaller.

Figure 1.9: Results of simulated rubella dynamics assuming constant birth rates no private sector vaccination. The figure shows the CRS incidence per 100,000 live births for rural Assam from 2016 to 2046 assuming constant birth rates as opposed to the figure above which allows declining birth rates. As above, the solid lines show the CRS incidence in the ‘routine + catch-up’ vaccine scenario for different assumed initial values of $R_0$. The dotted lines shows the CRS incidence over time in the ‘no vaccine’ scenario for different assumed initial values of $R_0$. The incidence of CRS is no vaccine scenarios is higher when birth rates are constant compared to falling birth rates. As a result, there is more risk of CRS risk ratio breaking one when birth rates remain constant as seem clearly when $R_0$ is nine.
1.4.2 Risk Factors for CRS Outbreaks

India currently has plans to introduce RCV into the public UIP (Sharan, 2014). Further evaluation of parameters such as $R_0$ and vaccination coverage are not currently being formally considered as part of the vaccine introduction program. Accordingly, in this section, we used our framework to evaluate the degree to which various sources of information, both from existing surveillance, but also other opportunistic efforts, could be used to identify risk factors which would enable health policy makers to discern whether vaccine introduction is having the desired effect of decreasing CRS incidence, or whether there is cause for concern. It is important to note that it was not tractable to explore the full range of stochastic models. The results are based on deterministic models and do not capture the randomness and uncertainty that inherently exists in all model parameters. Therefore, the following section proposes how data can be used operationally to make decisions following RCV introduction about the necessity of potential SIAs in regions. The following thresholds should not be interpreted as exact, despite detail in digits.

We focus on three possible negative outcomes of introduction of RCV into the public sector: i) the risk of an increase in cumulative 30-year CRS incidence; ii) the risk of transient annual increases in CRS incidence; and iii) the risk of any CRS outbreak (defined here as an incidence rate of $\geq 5$ cases per 100,000 live births). The first two undesired outcomes are necessarily assessed by comparing CRS incidence after the introduction of rubella into the public sector to CRS levels given that the vaccine was never introduced into the public sector. A 30-year increase in CRS burden implies a cumulative increase in the burden, and our results indicate that this is theoretically possible in some states under the assumption that $R_0$ is greater than nine (Figure 1.7); however, as discussed above such long-term cumulative effects have never been empirically observed in any country. Conversely, the second undesirable outcome, transient annual increases in CRS incidence, which may occur even if the cumulative burden has been reduced, have been reported in Greece (Panagiotopoulos et al., 1999) and Costa Rica (Morice et al., 2003; Metcalf et al., 2012b), and are therefore a potentially more immediate concern.
Table 1.1 shows a summary of risk factors for each of the three negative outcomes listed above, and the data that could be used to anticipate them. The risk factors are differentiated by dynamical, contextual, and conditional risk factors. A high $R_0$ of rubella is a dynamic risk factor, or an epidemiological parameter that directly determines rubella dynamics. Knowledge of $R_0$ is the gold standard, however it is often unknown and therefore we must rely on other measures that correlate with rubella dynamics and can capture increases in CRS. Insufficient routine rubella vaccination coverage is the second risk factor. Vaccination coverage is contextual in nature because it is determined programmatically. The third risk factor is a short honeymoon period. The honeymoon period is a period of low incidence after a mass vaccination campaign resulting from the reduced size of the susceptible population. The length of the honeymoon period is the result of $R_0$ (i.e., dynamical risk factor) and vaccination coverage (i.e., contextual risk factor) and is therefore described as a conditional risk factor.

**Dynamic risk factor for CRS outbreak: high rubella $R_0$**

A high rubella $R_0$ is a dynamic risk factor. A $R_0$ for rubella greater than nine may result in long-term CRS incidence ratio over one, and a $R_0$ over five may result in short-term transient increases in CRS incidence. As displayed in Figure 1.7 if the $R_0$ of rubella is nine or lower, the 30-year average CRS incidence ratio was less than one for all states. If the $R_0$ of rubella is five, transient increases in CRS incidence can be avoided for all states and regions (see [Winter (2016)] for link to online figure).

**Contextual risk factor for CRS outbreak: RCV coverage below critical threshold**

Rubella vaccination rates below a critical level of vaccination coverage represents a contextual risk factor and can result in unwanted outcomes. As discussed above, the critical level of vaccination coverage depends on population birth rate, transmissibility of disease, and whether vaccination has interrupted endemic transmission. Leveraging the diversity of demographic and vaccination
Table 1.1: Risk factors and needed sources of data for three undesired outcomes post introduction of RCV into the public sector in India: risk of average 30-year CRS incidence ratio (IR) > 1, risk of transient CRS incidence ratio (IR) > 1, and risk of any annual CRS incidence rate ≥ 5 cases per 100,000 live births. The results are based on the ‘routine + catch-up’ vaccine scenario.

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Risk Factors</th>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-year CRS IR &gt;1</td>
<td>$R_0 &gt; 9$</td>
<td>- age-stratified serological surveys</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- rubella age-incidence surveillance data</td>
</tr>
<tr>
<td></td>
<td>routine RCV vaccine coverage &lt; 0.77</td>
<td>- pre and post vaccination serological surveys</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- post vaccination serological testing that can distinguish vaccination and natural immunity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- vaccine coverage survey data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- vaccine administration data</td>
</tr>
<tr>
<td>transient CRS IR &gt;1</td>
<td>$R_0 &gt; 5$</td>
<td>- age-stratified serological surveys</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- rubella age-incidence surveillance data</td>
</tr>
<tr>
<td></td>
<td>routine RCV vaccine coverage &lt; 0.84</td>
<td>- pre and post vaccination serological surveys</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- post vaccination serological testing that can distinguish vaccination and natural immunity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- vaccine coverage survey data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- vaccine administration data</td>
</tr>
<tr>
<td>CRS incidence rate ≥ 5</td>
<td>routine RCV vaccine coverage &lt; 0.90</td>
<td>- pre and post vaccination serological surveys</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- post vaccination serological testing that can distinguish vaccination and natural immunity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- vaccine coverage survey data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- vaccine administration data</td>
</tr>
</tbody>
</table>

Coverage patterns observed across India, we characterize the minimum level of routine vaccine coverage required to avoid these outcomes for a range of values of $R_0$ (Table 1.2). Specifically, Table 1.2 lists the minimum level of coverage as observed by one of the 52 regions in which the following outcomes did not take place: an average 30-year CRS incidence ratio > 1, any transient or annual CRS incidence ratio > 1, and an outbreak of CRS of ≥ 5 cases per 100,000 live births. These results show that a lower vaccination coverage is needed to prevent undesired outcomes when a lower $R_0$ is assumed. The WHO’s critical vaccination threshold estimate of 80% routine immunization coverage will likely be high enough to prevent cumulative increase in CRS incidence for all values of $R_0$ and transient increases in CRS incidence as long as $R_0$ is less than or equal to nine.
Table 1.2: The minimum routine vaccination coverage level to avoid a 30-year average CRS incidence ratio (IR) > 1, to avoid any transient CRS incidence ratio (IR) > 1, and to avoid any annual CRS incidence rate ≥ 5 cases per 100,000 live births by assumed $R_0$ value. The results are based on the ‘routine + catch-up’ vaccine scenario.

<table>
<thead>
<tr>
<th>$R_0$</th>
<th>30-year CRS IR &gt; 1</th>
<th>transient CRS IR &gt; 1</th>
<th>CRS incidence rate ≥ 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>current levels sufficient</td>
<td>current levels sufficient</td>
<td>0.686</td>
</tr>
<tr>
<td>7</td>
<td>current levels sufficient</td>
<td>0.666</td>
<td>0.814</td>
</tr>
<tr>
<td>9</td>
<td>current levels sufficient</td>
<td>0.798</td>
<td>0.877</td>
</tr>
<tr>
<td>11</td>
<td>0.768</td>
<td>0.838</td>
<td>0.895</td>
</tr>
</tbody>
</table>

Conditional risk factor for CRS outbreak: short honeymoon period

The third risk factor, conditional in nature, is a short honeymoon period. We define the honeymoon period here as the number of months between introduction of the rubella vaccine in the public sector and a rubella outbreak including at least 5 cases per 100,000 person-years, or a CRS incidence rate of at least 5 per 100,000 live births. Tables 1.3 and 1.4 display the length of the honeymoon period for every region based on different assumed $R_0$ values. The length of the honeymoon period decreases as the assumed $R_0$ values decreases. We see that rubella outbreaks occur prior to CRS cases on average by at least ten months and provide an opportunity to administer SIAs that reach susceptible women of child bearing age before they are affected. Additionally, if the rubella honeymoon period is short, or less than eight years post vaccine introduction (2023) then it is likely that the $R_0$ for rubella is 11, and that states will experience both short and long-term CRS incidence increases.

Data sources of risk factors for CRS outbreak

For each undesired outcome, Table 1.1 also lists the data sources that can be used to evaluate the risk factors. Serological surveys provide a window into the landscape of population immunity for immunizing infections and thus provide invaluable information into the key variables here (Figure 1.10). If private sector vaccine is negligible, age-stratified serological data prior the vaccine
introduction directly corresponds to the cumulative proportion infected by each age \((\text{Grenfell and Anderson, 1985})\). Age stratified serological surveys of rubella antibodies prior to vaccine introduction can inform unknown epidemiological parameters of rubella such as \(R_0\) \((\text{Griffiths, 1974; Grenfell and Anderson, 1985; Cutts et al., 1999; Farrington et al., 2001; Gay et al., 1995; Chakraborty et al., 1973})\). Serological surveys before and after vaccine campaigns are ideal for evaluating vaccination coverage levels and the success of a vaccination program \((\text{Andrews et al., 2008})\). However, in the event that future serological testing can distinguish IgG antibodies as either from natural immunization or vaccination, only post vaccination serological surveys would be necessary. Given the existence of rubella vaccination in the private sector (and public sector for Delhi, Goa, Puducherry, and Sikkim), it is necessary that any serological survey also include a questionnaire that asks about past rubella vaccinations. Estimation of \(R_0\) will be overestimated if childhood vaccinations are not taken into account.

![Figure 1.10](image)

Figure 1.10: Results of simulated rubella dynamics in the ‘routine + catch-up’ vaccine scenario, assuming no private sector vaccination. a) 2015 pre-RCV introduction proportion of rubella seropositive individuals by age and \(R_0\). b) 2017 1 year post-RCV introduction proportion of rubella seropositive individuals by age and \(R_0\).
Rubella age-incidence data is another important source of information. Cross-sectional age-incidence data can be used to roughly estimate $R_0$ based on the average age of infection and mean life expectancy (Anderson and May, 1991). Age-incidence data over time can be used in TSIR models to more accurately estimate $R_0$ (Bjornstad et al., 2002; Finkenstadt and Grenfell, 2000; Metcalf et al., 2009; Farrington et al., 2001). Rubella incidence data over time additionally identifies rubella outbreaks and can be used to estimate the average number of months until a rubella outbreak post-vaccination introduction. Unfortunately, there is currently no reliable national surveillance and registry for rubella or CRS in India. Rubella incidence surveillance is a case-based surveillance system, and depends upon laboratory of epidemiologically confirmed cases from individuals who present for care with a rash and fever. Rubella cases are believed to be highly under-reported in India (Dewan and Gupta, 2012; Strebel et al., 2014), similar to other middle and low income countries (Metcalf et al., 2011b,a; Strebel et al., 2014). Tables 1.3 and 1.4 also display the required reporting rate to capture the rubella outbreak of $\geq 5$ rubella cases per 100,000 person-years and the CRS outbreak of $\geq 5$ CRS cases per 100,000 live births. Surveillance would need to achieve upwards of a 75% reporting rate to capture most post-honeymoon period rubella outbreaks, and 65% to capture most post-honeymoon period CRS outbreaks.

Vaccination coverage survey data and vaccine administration data are readily available data sources to estimate vaccination coverage rates (ICF International). These sources of data however are prone to reporting bias of vaccination coverage as compared to pre and post vaccination campaign serological surveys (Lessler et al., 2011a).
Table 1.3: The number of months until rubella outbreak (5 rubella cases per 100,000 population) post-vaccine introduction, by $R_0$ and for each state by region. The table also displays the rubella cases reporting rate required to pick up an outbreak of $\geq 5$ per 100,000 population. The results are based on the ‘routine + catch-up’ vaccine scenario.

<table>
<thead>
<tr>
<th>State Regions</th>
<th>$R_0=5$</th>
<th>$R_0=7$</th>
<th>$R_0=9$</th>
<th>$R_0=11$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andhra Pradesh Rural</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Andhra Pradesh Urban</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Arunachal Pradesh Rural</td>
<td>116 (0.23)</td>
<td>79 (0.19)</td>
<td>65 (0.07)</td>
<td>55 (0.19)</td>
</tr>
<tr>
<td>Arunachal Pradesh Urban</td>
<td>158 (0.88)</td>
<td>103 (0.13)</td>
<td>79 (0.17)</td>
<td>67 (0.17)</td>
</tr>
<tr>
<td>Assam Rural</td>
<td>-</td>
<td>104 (0.38)</td>
<td>79 (0.24)</td>
<td>67 (0.2)</td>
</tr>
<tr>
<td>Assam Urban</td>
<td>-</td>
<td>133 (0.81)</td>
<td>93 (0.51)</td>
<td>78 (0.2)</td>
</tr>
<tr>
<td>Bihar Rural</td>
<td>-</td>
<td>82 (0.92)</td>
<td>67 (0.27)</td>
<td>57 (0.56)</td>
</tr>
<tr>
<td>Bihar Urban</td>
<td>-</td>
<td>104 (0.42)</td>
<td>79 (0.26)</td>
<td>67 (0.21)</td>
</tr>
<tr>
<td>Chhattisgarh Rural</td>
<td>-</td>
<td>-</td>
<td>91 (0.34)</td>
<td>77 (0.13)</td>
</tr>
<tr>
<td>Chhattisgarh Urban</td>
<td>-</td>
<td>-</td>
<td>94 (0.88)</td>
<td></td>
</tr>
<tr>
<td>Gujarat Rural</td>
<td>-</td>
<td>-</td>
<td>107 (0.47)</td>
<td>90 (0.21)</td>
</tr>
<tr>
<td>Gujarat Urban</td>
<td>-</td>
<td>-</td>
<td>93 (0.57)</td>
<td>78 (0.21)</td>
</tr>
<tr>
<td>Haryana Rural</td>
<td>-</td>
<td>-</td>
<td>95 (0.3)</td>
<td>80 (0.38)</td>
</tr>
<tr>
<td>Haryana Urban</td>
<td>-</td>
<td>-</td>
<td>120 (0.78)</td>
<td>92 (0.51)</td>
</tr>
<tr>
<td>Himachal Pradesh Rural</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Himachal Pradesh Urban</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jammu &amp; Kashmir Rural</td>
<td>-</td>
<td>-</td>
<td>104 (0.34)</td>
<td></td>
</tr>
<tr>
<td>Jammu &amp; Kashmir Urban</td>
<td>-</td>
<td>-</td>
<td>105 (0.51)</td>
<td></td>
</tr>
<tr>
<td>Jharkhand Rural</td>
<td>93 (0.5)</td>
<td>75 (0.87)</td>
<td>64 (0.05)</td>
<td></td>
</tr>
<tr>
<td>Jharkhand Urban</td>
<td>-</td>
<td>105 (0.52)</td>
<td>81 (0.66)</td>
<td></td>
</tr>
<tr>
<td>Karnataka Rural</td>
<td>-</td>
<td>-</td>
<td>119 (0.93)</td>
<td></td>
</tr>
<tr>
<td>Karnataka Urban</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kerala Rural</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>117 (0.65)</td>
</tr>
<tr>
<td>Kerala Urban</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>104 (0.4)</td>
</tr>
<tr>
<td>Madhya Pradesh Rural</td>
<td>-</td>
<td>91 (0.3)</td>
<td>69 (0.67)</td>
<td>63 (0.04)</td>
</tr>
<tr>
<td>Madhya Pradesh Urban</td>
<td>-</td>
<td>-</td>
<td>119 (0.79)</td>
<td>91 (0.53)</td>
</tr>
<tr>
<td>Maharashtra Rural</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Maharashtra Urban</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Manipur Rural</td>
<td>108 (0.21)</td>
<td>83 (0.82)</td>
<td>71 (0.71)</td>
<td></td>
</tr>
<tr>
<td>Manipur Urban</td>
<td>133 (0.8)</td>
<td>93 (0.55)</td>
<td>78 (0.22)</td>
<td></td>
</tr>
<tr>
<td>Meghalaya Rural</td>
<td>120 (0.74)</td>
<td>90 (0.2)</td>
<td>76 (0.95)</td>
<td></td>
</tr>
<tr>
<td>Meghalaya Urban</td>
<td>130 (0.52)</td>
<td>92 (0.3)</td>
<td>78 (0.12)</td>
<td></td>
</tr>
<tr>
<td>Mizoram Rural</td>
<td>117 (0.55)</td>
<td>82 (0.81)</td>
<td>69 (0.68)</td>
<td></td>
</tr>
<tr>
<td>Mizoram Urban</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nagaland Rural</td>
<td>159 (0.63)</td>
<td>95 (0.87)</td>
<td>78 (0.12)</td>
<td>66 (0.12)</td>
</tr>
<tr>
<td>Nagaland Urban</td>
<td>289 (0.88)</td>
<td>107 (0.75)</td>
<td>82 (0.64)</td>
<td>70 (0.58)</td>
</tr>
<tr>
<td>Odisha Rural</td>
<td>-</td>
<td>104 (0.46)</td>
<td>81 (0.64)</td>
<td></td>
</tr>
<tr>
<td>Odisha Urban</td>
<td>-</td>
<td>-</td>
<td>105 (0.58)</td>
<td></td>
</tr>
<tr>
<td>Punjab Rural</td>
<td>-</td>
<td>-</td>
<td>106 (0.82)</td>
<td></td>
</tr>
<tr>
<td>Punjab Urban</td>
<td>-</td>
<td>-</td>
<td>119 (0.92)</td>
<td></td>
</tr>
<tr>
<td>Rajasthan Rural</td>
<td>94 (0.76)</td>
<td>77 (0.1)</td>
<td>65 (0.09)</td>
<td></td>
</tr>
<tr>
<td>Rajasthan Urban</td>
<td>-</td>
<td>96 (0.35)</td>
<td>80 (0.44)</td>
<td></td>
</tr>
<tr>
<td>Tamil Nadu Rural</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tamil Nadu Urban</td>
<td>-</td>
<td>-</td>
<td>119 (0.9)</td>
<td></td>
</tr>
<tr>
<td>Tripura Rural</td>
<td>-</td>
<td>107 (0.33)</td>
<td>90 (0.19)</td>
<td></td>
</tr>
<tr>
<td>Tripura Urban</td>
<td>-</td>
<td>-</td>
<td>107 (0.39)</td>
<td></td>
</tr>
<tr>
<td>Uttar Pradesh Rural</td>
<td>95 (0.91)</td>
<td>69 (0.52)</td>
<td>57 (0.46)</td>
<td>52 (0.02)</td>
</tr>
<tr>
<td>Uttar Pradesh Urban</td>
<td>117 (0.37)</td>
<td>79 (0.22)</td>
<td>65 (0.08)</td>
<td>53 (0.2)</td>
</tr>
<tr>
<td>Uttarakhand Rural</td>
<td>-</td>
<td>93 (0.7)</td>
<td>78 (0.18)</td>
<td></td>
</tr>
<tr>
<td>Uttarakhand Urban</td>
<td>-</td>
<td>121 (0.7)</td>
<td>92 (0.43)</td>
<td></td>
</tr>
<tr>
<td>West Bengal Rural</td>
<td>-</td>
<td>-</td>
<td>105 (0.59)</td>
<td></td>
</tr>
<tr>
<td>West Bengal Urban</td>
<td>-</td>
<td>-</td>
<td>106 (0.71)</td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>137.5 (0.755)</td>
<td>103.5 (0.52)</td>
<td>92 (0.46)</td>
<td>81 (0.4)</td>
</tr>
</tbody>
</table>
Table 1.4: The number of months until CRS incidence rate of 5 cases per 100,000 live births post-vaccine introduction, by $R_0$ and for each state by region. The table also displays the CRS cases reporting rate required to pick up an incidence rate of $\geq 5$ cases per 100,000 live births. The results are based on the ‘routine + catch-up’ vaccine scenario.

<table>
<thead>
<tr>
<th>State Regions</th>
<th># months until CRS incidence rate $\geq 5$ (reporting rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_0=5$</td>
</tr>
<tr>
<td>Andhra Pradesh Rural</td>
<td>-</td>
</tr>
<tr>
<td>Andhra Pradesh Urban</td>
<td>-</td>
</tr>
<tr>
<td>Arunachal Pradesh Rural</td>
<td>130 (0.51) 91 (0.25) 76 (0.19) 66 (0.3)</td>
</tr>
<tr>
<td>Arunachal Pradesh Urban</td>
<td>181 (0.76) 114 (0.17) 89 (0.15) 77 (0.16)</td>
</tr>
<tr>
<td>Assam Rural</td>
<td>-</td>
</tr>
<tr>
<td>Assam Urban</td>
<td>-</td>
</tr>
<tr>
<td>Bihar Rural</td>
<td>-</td>
</tr>
<tr>
<td>Bihar Urban</td>
<td>-</td>
</tr>
<tr>
<td>Chhattisgarh Rural</td>
<td>-</td>
</tr>
<tr>
<td>Chhattisgarh Urban</td>
<td>-</td>
</tr>
<tr>
<td>Gujarat Rural</td>
<td>-</td>
</tr>
<tr>
<td>Gujarat Urban</td>
<td>-</td>
</tr>
<tr>
<td>Haryana Rural</td>
<td>-</td>
</tr>
<tr>
<td>Haryana Urban</td>
<td>-</td>
</tr>
<tr>
<td>Himachal Pradesh Rural</td>
<td>-</td>
</tr>
<tr>
<td>Himachal Pradesh Urban</td>
<td>-</td>
</tr>
<tr>
<td>Jammu &amp; Kashmir Rural</td>
<td>-</td>
</tr>
<tr>
<td>Jammu &amp; Kashmir Urban</td>
<td>-</td>
</tr>
<tr>
<td>Jharkhand Rural</td>
<td>-</td>
</tr>
<tr>
<td>Jharkhand Urban</td>
<td>-</td>
</tr>
<tr>
<td>Karnataka Rural</td>
<td>-</td>
</tr>
<tr>
<td>Karnataka Urban</td>
<td>-</td>
</tr>
<tr>
<td>Kerala Rural</td>
<td>-</td>
</tr>
<tr>
<td>Kerala Urban</td>
<td>-</td>
</tr>
<tr>
<td>Madhya Pradesh Rural</td>
<td>-</td>
</tr>
<tr>
<td>Madhya Pradesh Urban</td>
<td>-</td>
</tr>
<tr>
<td>Maharashtra Rural</td>
<td>-</td>
</tr>
<tr>
<td>Maharashtra Urban</td>
<td>-</td>
</tr>
<tr>
<td>Manipur Rural</td>
<td>-</td>
</tr>
<tr>
<td>Manipur Urban</td>
<td>-</td>
</tr>
<tr>
<td>Meghalaya Rural</td>
<td>-</td>
</tr>
<tr>
<td>Meghalaya Urban</td>
<td>-</td>
</tr>
<tr>
<td>Mizoram Rural</td>
<td>-</td>
</tr>
<tr>
<td>Mizoram Urban</td>
<td>-</td>
</tr>
<tr>
<td>Nagaland Rural</td>
<td>196 (0.88) 114 (0.88) 89 (0.16) 77 (0.18)</td>
</tr>
<tr>
<td>Nagaland Urban</td>
<td>364 (0.99) 128 (0.29) 95 (0.54) 80 (0.38)</td>
</tr>
<tr>
<td>Odisha Rural</td>
<td>-</td>
</tr>
<tr>
<td>Odisha Urban</td>
<td>-</td>
</tr>
<tr>
<td>Punjab Rural</td>
<td>-</td>
</tr>
<tr>
<td>Punjab Urban</td>
<td>-</td>
</tr>
<tr>
<td>Rajasthan Rural</td>
<td>-</td>
</tr>
<tr>
<td>Rajasthan Urban</td>
<td>-</td>
</tr>
<tr>
<td>Tamil Nadu Rural</td>
<td>-</td>
</tr>
<tr>
<td>Tamil Nadu Urban</td>
<td>-</td>
</tr>
<tr>
<td>Tripura Rural</td>
<td>-</td>
</tr>
<tr>
<td>Tripura Urban</td>
<td>-</td>
</tr>
<tr>
<td>Uttar Pradesh Rural</td>
<td>116 (0.31) 87 (0.55) 72 (0.46) 63 (0.43)</td>
</tr>
<tr>
<td>Uttar Pradesh Urban</td>
<td>142 (0.48) 93 (0.42) 77 (0.26) 68 (0.41)</td>
</tr>
<tr>
<td>Uttarakhand Rural</td>
<td>-</td>
</tr>
<tr>
<td>Uttarakhand Urban</td>
<td>-</td>
</tr>
<tr>
<td>West Bengal Rural</td>
<td>-</td>
</tr>
<tr>
<td>West Bengal Urban</td>
<td>-</td>
</tr>
<tr>
<td>Average</td>
<td>161.5 (0.635) 116 (0.48) 101 (0.65) 93 (0.62)</td>
</tr>
</tbody>
</table>
1.5 Discussion

There is increasing recognition of the importance of moving beyond country-scale analyses to consider heterogeneity within countries when evaluating public health interventions (Ferrari et al., 2013). States across India differ in terms of demographic rates such as population size, births, deaths, and population age structure, and in terms of health indicators such as vaccination coverage rates, and all of these differences will affect the outcome of public health efforts. Here, we provide a detailed state by state analysis of the introduction of rubella containing vaccine into India’s public sector childhood UIP.

Our results indicate that, given India’s demography, introduction of RCV into the India’s public sector immunization program will likely be broadly beneficial for all simulated states, given one key condition. Empirical analysis estimated that the median $R_0$ across the continent of Africa was 5.2 (90% CI 4.0-6.7) (Lessler and Metcalf, 2013). If we assume that the $R_0$ of rubella across Africa is similar across India, then our analysis confirms that the introduction of the rubella vaccine into the public sector across India will result in a decrease in the long and short-term incidence of CRS in all states compared to if the vaccine was never introduced. However, if the $R_0$ for rubella is seven or higher as seen in Addis Ababa, Ethiopia (Cutts et al., 2000), then introduction of the rubella vaccine into the public sector can result in short-term increases in CRS incidence rates in regions where routine vaccination coverage rates are below 84%.

Vynnycky et al. (2003) demonstrated that private sector rubella vaccination at levels found in India (<60%) would lead to increases in the incidence of CRS both among unvaccinated individuals and the general population. Our sensitivity analysis on private sector vaccinations also showed that introduction of the rubella vaccine into the public sector will help drive down CRS incidence even more in the event of private sector vaccinations.
Considering India’s upcoming plan to roll-out RCV (British Broadcasting Company 2014; Hindustan Times 2014), we also highlight potential risk factors that health officials can use to determine regions where the incidence of CRS may potentially increase after the introduction of the vaccine: high $R_0$, low vaccination coverage, and short length of the honeymoon period. An improved case-based surveillance system and/or rubella serological surveys are crucial to be able to identify these risk factors. The reporting rate for rubella cases must be between 65 and 75% to capture most post-honeymoon period rubella outbreaks. No population-based rubella serological survey has been conducted in India (Dewan and Gupta 2012). However, rubella serological surveys are pertinent to understanding the landscape of rubella infections prior to vaccine introduction, informing country specific knowledge gaps of disease burden and unknown transmission dynamics, evaluating the success of vaccination programs, and determining if SIAs are necessary and if so which age groups to target (Cutts et al. 1999; Andrews et al. 2008; Babad et al. 1995; Ximenes et al. 2014; Chakraborty et al. 1973). Evaluation of multiple pathogens simultaneously is a cost-effective option to administering serological surveys for rubella (World Health Organization 2011a,b).

While our model extends previous work to consider the details of within-country heterogeneity, our analysis remained deterministic, assumed a well-mixed spatial population, and did not evaluate the impact of timing of vaccination or likely extinction recolonization dynamics. Aron (1990) theoretically demonstrated that the timing of vaccinations affects the stable equilibrium of an epidemic cycle; as a result, the length of the honeymoon period post vaccination introduction would be influenced by the timing of RCV introduction. Additionally, previous work has indicated that extinction and recolonization dynamics of rubella can shape the burden of CRS (Metcalf et al. 2011b; Clarke et al. 1980). This limitation is of particular concern in the simulations where $R_0$ was assumed to be five, because such relatively low values of $R_0$ mean that local extinction of the infection is likely, and large pools of susceptible women of reproductive age might then build up, both in the absence of vaccination and in scenarios where vaccination is occurring. Where this is the case, while our conclusions as to the country and state-scale benefits of introduction of RCV are likely to hold, a subset of local communities, particularly those where
vaccination coverage is low may even experience an increase in the burden of CRS following introduction of RCV as women are infected following reintroduction of the virus (Metcalf et al., 2013). Previous work shows that vaccination coverage is associated with spatially correlated determinants such as mother’s education, household wealth, and health facility infrastructure (International Institute for Population Sciences and Macro International, 2007; Datar et al., 2007); incorporating such details into a similar framework represent an interesting direction for future work to determine when and how inequitable outcomes of introduction of RCV might be of concern.

The reliability of model based simulations are only as good as parameter estimates. We assumed seasonal forcing of rubella infections, although the degree of forcing has a negligible affect on CRS estimates (Metcalf et al., 2012a). We assumed that the rate of contact by age for India was similar to Great Britain based in the European POLYMOD diary study (Mossong et al., 2008). Analyses in other countries indicate qualitatively similar patterns (Horby et al., 2011; Mossong et al., 2008; DeStefano et al., 2011), so biases may not be too considerable, but ultimately, data from the focal populations would strengthen the analysis. Estimated vaccination coverage rates were based on a sophisticated small area estimation model that analyzed five different population-based surveys; however it is still prone to information bias based on mother’s recall, and the accuracy of vaccination cards (Pramanik et al., 2015). We assumed that rubella vaccine coverage rates would be equal to 2011 measles dose one vaccine coverage rates, and remained constant over time. Given the introduction of measles containing vaccine dose two to India’s UIP, the global focus on eliminating measles, and the likelihood that most future measles containing vaccine administered to children will be MR vaccine, our rubella vaccine coverage rates likely underestimate future rates. As such, the model makes conservative predictions of the benefits of RCV introduction, and CRS incidence may be overestimated post vaccine introduction. Note, these simulations also do not take into account additional behavioral changes over time. The country’s policies or individual’s preference that influence birth, death, migration, vaccination coverage rates, and rate of contact across ages will likely change over time. Despite these assumptions and caveats, disease dynamic projections are important analytical tools for policy makers.
Our results provide optimism for India’s plan to add rubella containing vaccine to their public UIP (British Broadcasting Company, 2014; Hindustan Times, 2014). For all states’ rural and urban regions, we found that the introduction of RCV will result in decreases in cumulative CRS incidence, assuming an $R_0$ of nine or below. In the case of higher $R_0$, or in concern over transient increases in CRS, specified risk factors above can be use by health program officers and policy makers as guidelines to identify potential regions at risk of CRS incidence increases. As such, India’s readiness and flexibility to administer supplementary immunization activities (SIA), as was needed in Costa Rica, will be valuable if these risk factors appear (Morice et al., 2003; Metcalf et al., 2012b). Future research on RCV in India’s UIP will need to address spatial movement between states and regions within states, and explore a variety of spatial roll-out scenarios across the country.
Towards Measles Elimination: Exploring the Potential of Nested Serological Surveys within Measles and Rubella Surveillance Systems to Identify the Accumulation of Measles Susceptible Individuals in Madagascar

2.1 Abstract

*Background:* Despite high measles vaccination coverage rates, large measles outbreaks have occurred in some sub-Saharan African countries due to an unobserved accumulation of susceptible individuals. Current measles and rubella surveillance systems, which focus solely on incidence data via fever-rash, and IgM serological testing, are at risk of not recognizing an accumulation of susceptible individuals because of dwindling measles incidence data. Serological surveys that test for measles IgG antibodies can fill this gap in information by estimating seroprevalence by age to inform vaccination strategies, prevent future outbreaks, and eliminate measles. This paper explores the use of nested measles IgG serological surveys as a low cost extension to the measles and rubella surveillance system in Madagascar. *Methods:* We analyzed 4394 serum samples that were collected between November 2004 and December 2015 from the measles and rubella surveillance system in Madagascar. Using semiparametric models with penalized regression smoothers, we empirically
estimated the current proportion of seropositive individuals by age. To evaluate the strength of this inference, we compared direct estimates of seroprevalence by age to indirect estimates based on a birth cohort projection of immunity. We also explored the effect of different Supplementary Immunization Activities (SIAs) to prevent a measles outbreak and move towards measles elimination. 

Results: Based on the serological data, we estimated that 84.9% of the total Malagasy population is immune to measles (95% CI: 75.3%, 89.8%). The indirect estimate falls just inside the 95% confidence bounds of this point estimate; however, the two age profiles of seroprevalence differed markedly over age. Our results indicate that Madagascar is at risk of a major measles outbreak. According to the data, past Madagascar measles SIAs that have targeted children up to five years old will not be sufficient to effectively reduce the risk of a major measles outbreak; rather, SIAs that target individuals up to 15 or 20 years old would be more effective at reducing the number of susceptible individuals. Conclusions: As a low-cost alternative to population-based serological surveys, surveillance system nested measles IgG serological surveys have potential as a measles control optimization tool in settings of dwindling incidence data. This analysis took the first step in evaluating this data source by comparing estimates to birth cohort projection of immunity analysis indirect estimates. Further replication by comparison to population or community-based serological data is recommended before inclusion into existing surveillance is considered. Despite residual uncertainty in the age profile of immunity estimated, both direct and indirect estimates indicated that Malagasy population immunity was less than 90%. As a result, we recommend a high impact measles SIA, at least up to 15 years old, to reduce the risk of a large measles outbreak and make strides toward elimination.

### 2.2 Introduction

Measles virus is highly infectious and is transmitted via respiratory droplets and direct contact (World Health Organization, 2009). Measles symptoms, which generally appear 10 to 12 days after infection, include high fever, rash, cough, and conjunctivitis (World Health Organization, 2009). Despite the successful reduction in measles deaths by 75% since 2000, measles remains a significant cause of morbidity and mortality in the world. Twenty million individuals are infected with measles annually, and an estimated 115,000 measles deaths occurred globally in 2014 (World Health Organization, 2016a).

The measles vaccine is safe, effective, and inexpensive (Reef and Plotkin, 2013). All six regions of the WHO now have measles elimination targets (CDC, 2014). To eliminate measles, vaccination programs must sustain high vaccination coverage rates (about 95%) that interrupt measles transmission in the population and reach herd immunity (World Health Organization, 2009). In
the event that countries are unable to achieve and maintain high vaccination coverage via routine immunization and Supplementary Immunization Activities (SIAs), susceptible individuals may accumulate over time and there is a risk of a large outbreak of measles post vaccine introduction (McLean 1995; World Health Organization 2011b).

The WHO defined region of the Americas was able to achieve and maintain a high vaccination coverage and successfully eliminated endemic measles from the region as of November 2002 (World Health Organization 2009). This process included: vaccinating over 90% of each birth cohort and increasing the age of first dose up to 12 months; catch-up campaigns targeting children up to 15 years old; and supplementary immunization activities (SIAs) vaccinating children up to 5 years old, and up to age 40 years old once the rubella vaccine was added to the measles vaccine (Cutts et al., 2013). However, the Americas’ measles control strategy can not simply be exported to other regions because factors that affect measles dynamics, such as access to care, childhood nutrition, HIV rates, maternal immunity (Mckee et al., 2015) and demographic rates (Ferrari et al., 2013), are highly variable across regions. Control efforts must thus be targeted to the region of interest for maximum efficacy.

In fact, even countries thought to have successful vaccination coverage rates have experienced large outbreaks due to a hidden accumulation of susceptible individuals, such as occurred in Malawi and Burkina Faso (Chen et al., 1994; Cutts et al., 2013; Kidd et al., 2012; Minetti et al., 2013b,a; Shibeshi et al., 2014). Malawi experienced a long period of low measles incidence where susceptible individuals unknowingly built up at older ages, resulting in a large outbreak of 134,000 cases and 304 deaths in 2010 (Minetti et al., 2013b; Cutts et al., 2013). Current measles and rubella surveillance systems, which focus solely on incidence data, are at risk of not recognizing an accumulation of susceptible individuals because of dwindling incidence data. Dwindling incidence data is caused by an actual reduction in incidence cases due to vaccination, potentially combined with low reporting rates. As a result, case notification data becomes an ambiguous predictor of outbreak potential. Therefore, in settings of dwindling incidence data, researchers must rely solely
on estimated vaccination coverage to predict future measles outbreaks and advise of the need for SIAs (Takahashi et al., 2015). Further, in many countries, there is often no direct measure of the potentially complex profile of immunity over age resulting from past vaccination efforts and measles outbreaks, which could aid in the targeting of SIAs.

Serological surveys that test for measles Immunoglobulin G (IgG) antibodies are an under-utilized tool that can fill this gap in information and assess the potential of a susceptible build-up. Measles IgG antibodies are a marker of individual measles immunity, the result of either natural infection or vaccination. IgG serological surveys, therefore, estimate population immunity profiles of measles by age, and as a result can identify age groups to be included in SIAs and evaluate the effectiveness of previous vaccine control efforts at reducing population susceptibility (Cutts and Hanson, 2016; Gay et al., 1997; Nigatu et al., 2008). For example, measles IgG serological testing in England and Wales in 1994 contributed to a decision to implement a catch-up measles campaign among school-aged children (Babad et al., 1995), which is believed to have avoided a measles outbreak (Osborne et al., 1997).

Our analysis is based on the use of existing serum samples from the measles and rubella fever-rash surveillance system in Madagascar. Typically, such samples, which exist in a number of countries, are only tested for measles and rubella Immunoglobulin M (IgM) antibodies, a short term indicator of infection. Here, we leveraged existing banked samples to also test for IgG antibodies, a marker of immunity. To the best of our knowledge, the potential for such a low-cost extension of existing surveillance data to guide elimination and control efforts by directly revealing the age profile of immunity has never been systematically explored previously. Ideally, resulting estimates would be evaluated in the face of population based serological surveys. In the absence of such data, we evaluated our estimates by comparing them to recently developed indirect estimates of population immunity (Takahashi et al., 2015). To explore the implications of our analysis for investments of control, we used a direct probabilistic estimate of age profile of immunity across Madagascar to highlight the short-term successes of measles elimination programs, and to explore how well-
designed SIAs, which target gaps in immunity by age, can prevent the potentially looming measles outbreak in Madagascar.

2.3 Methods

2.3.1 Data

Serum samples were obtained from the Malagasy system of general surveillance for measles and rubella that were collected between November 2004 and December 2015. Ideally, when patients present for care with signs consistent with clinical criteria for measles (fever and rash and either cough, coryza, or conjunctivitis), the patients’ serums are collected and subsequently tested for measles IgM antibodies using standard serological techniques at the WHO national reference laboratory located at the Institut Pasteur de Madagascar (IPM) in Antananarivo. The unique feature of this data was that patients’ serum was also tested for measles IgG antibodies. Samples positive for measles IgG antibodies were considered seropositive and immune to measles, otherwise samples were considered seronegative or susceptible to measles. Samples considered IgG antibody equivocal were retested; if they remained equivocal (235 samples) we assumed IgG seropositivity. Due to misclassification bias and the potential for waning immunity, seroprevalence is not exactly equal to measles immunity, although for this analysis we assumed that the two correspond exactly.

A total of 4584 serum samples were collected from patients between November 2004 and December 2015. One-hundred and eighteen samples did not contain enough serum to test for measles IgG antibodies, and were removed from the analysis. Patient observations were additionally dropped if age was unknown (i.e., either age in years was not recorded or a combination of date of birth and date the specimen was collected was not recorded; 17 samples), or if the patient was older than 65 years of age (1 sample). We also removed one outlier sample observation of an individual that tested seronegative at age 42 years old. Between 2004 and 2015, 52 samples tested positive
for measles IgM antibodies (42 of which also tested positive for measles IgG antibodies). We intended to explore the potential for the measles surveillance system sample to provide a generalizable sample of population measles immunity; therefore, we decided to remove these 52 samples from the main analysis to minimize oversampling of measles seropositive individuals (albeit from natural infection). If these samples were left in the final analysis, we would need to code the 10 IgM seropositive and IgG seronegative as IgG seropositive, because they would gain immunity from their natural infection, which would oversample measles immunity from natural infection.

The total sample size that remained after removals was 4394 samples. If the patients’ age in months was not specified, we randomly assigned a birth month based on a uniform distribution. Additional information was recorded, including sex of the patient and the province, region, and district from which the sample was collected.

2.3.2 Data Based on Non-Probability Sample

Surveillance data is a non-probability sample, meaning it is not generated by random selection of individuals into the sample, and thus may not be representative of the Malagasy population. Figure 2.1 shows the population age structure of the sample compared to the 2015 population age structure per United Nations Population Division (UNPD) estimates for Madagascar (United Nations, 2015). Younger ages (less than 15 years old) were oversampled in every year and cumulatively across all years. We therefore corrected for this in the analysis (see below).

The sex ratio of the sampled population was 2070:2312 (males to females), which means that the surveillance system oversampled females. Measles infection is not associated with gender (Anderson and May, 1991); therefore, the potential for bias based on oversampling females depends on the association between sex and measles vaccination. Studies from Africa have shown null and negative associations between being male and receiving vaccine, which either means there was no bias, or we
Figure 2.1: Population and sample age structure by five-year age groups. The sample population based on the surveillance data is in colored bars from 2004 to 2015. The general population distribution based on UNPD data is in hollow grey outlined bars (United Nations 2015).

were overestimating immunity from vaccination (Antai 2012; Canavan et al. 2014; Lakew et al. 2015).

We also know the province, region, and district from where each sample was sent. Figure 2.2A shows the ratio of the observed proportion of samples per region to the expected proportion of samples per region based on the proportion of the total Malagasy population residing in each region (Institut National de la Statistique Madagascar 2004). It shows that the northwest region of Diana, and the central east region of Alaotra Mangoro were the most oversampled. Figure 2.2B gives us some idea of the result of this non-representative sampling by region in terms of measles seropositivity from vaccination. The most oversampled region of Alaotra Mangoro had higher vaccination coverage rates (at least according to administrative estimates) compared to the country mean vaccination coverage. Vaccination coverage in the northwest region of Diana was approximately equal to the country average measles vaccination coverage, so oversampling of this region did not bias the sample in terms of measles seropositivity from vaccination. The southwest region of Atsimo-Andrefana was oversampled, however it had a lower vaccination coverage compared to the country average vaccination coverage. The mean vaccination coverage ratio weighted by the sampling ratio by region was 0.99, which means that on average, the sample
Figure 2.2: Sampling and vaccination ratios region comparison of sampling ratio of observed to expected samples, and vaccination ratio of observed to mean measles vaccination coverage. A) sampling ratio of observed to expected samples, specifically it is the proportion of the samples from each region divided by the proportion of Madagascar’s population that resides in each region (Institut National de la Statistique Madagascar, 2004). B) vaccination ratio of observed to mean measles vaccination coverage, specifically it is the ratio of measles vaccination coverage by region divided by the average measles vaccination coverage, which is 0.69 (Institut National de la Statistique Madagascar and ICF Macro, 2010; World Health Organization and UNICEF, 2015) very slightly underestimated the mean proportion vaccinated across the country. Therefore, the sampling bias by region effectively had a null effect on measles immunity via measles vaccination.

The relationship between sampling bias by region and measles immunity via natural infection cannot be empirically assessed given the lack of measles incidence data. However, we would expect remote regions to have lower measles seropositivity via natural infection because remote regions have also been shown to have lower rubella incidence (Wesolowski et al., 2016), and spatial transmission patterns for these two infections are likely similar. If we assumed that rubella infections are the main drivers of individuals who present for care with fever and rash in Madagascar, then sampling by region may not be biased for measles seropositivity via natural infection.
2.3.3 Direct Empirical Estimates of Proportion Seropositive by Age - Semi-parametric Approach

We directly estimated the proportion of seropositive individuals based on the IgG serosurvey data. Specifically, we estimated: i) the total proportion of seropositive individuals in the population for each complete year of data (2005 to 2015), ii) the proportion of seropositive individuals under the age of five years old for each complete year of data (2005 to 2015), and iii) the current age profile of seropositivity.

To estimate the proportion seropositive under five years old and the total population proportion seropositive by year, we first smoothed the data over age and by year, and then extracted age-weighted average proportions seropositive based on Madagascar’s population age structure (United Nations, 2015). We chose to use semiparametric models with penalized regression smoothers to estimate measles seropositivity by age and year, based on the ‘mgcv’ library in R (Wood, 2006, 2011). Semiparametric approaches are ideal for assessing seroprevalence by age in non-endemic settings, and also ideal when using non-probability samples because semiparametric approaches are flexible and can ignore the constraint of monotonicity across age that is assumed by parametric models (Hens et al., 2012). We chose the final penalized regression smoothers model based on the lowest AIC, BIC, and residual deviance (see Appendix B.1.1 for details on model choice). The final model included isotropic smoothers on age and year with a cubic spline base, and a tensor product smoother on the interaction term between age and year, also with a cubic spline base (Wood, 2006, 2011). We used a binomial probability distribution and a logistic link function. The selected model estimated an age profile of immunity by year, which we would expect, given that population demography and measles dynamics are changing over time. Given the limited number of samples by age and per year, the lower bound of the smoothed age profile dropped to zero. To correct for this implausible outcome, we assumed that the minimum proportion seropositive was 0.80 after the age of 5 years old. For more information on our model choice used to estimate seropositivity by year, see Appendix B.1.1.
Despite the fact that the model selected to obtain yearly estimates provides age profile of immunity curve for each year, including the most recent 2015, we decided not to rely on this model for two reasons: i) unexpected changes in estimated proportion seropositive by age across years, ii) the sample size of 2015 data, specifically among 15 years old is considerably smaller than the previous two years (see Appendix B.1.2 for more detail). Therefore, to estimate the current age profile of immunity, we chose to combine all samples collected between November 2013 and December 2015 (1083 samples) to increase the sample size and potentially the generalizability of the sample. Second, we smoothed the data over age using penalized regression smoothers. Combining data across years increased the sample size and representativeness of the data; additionally, we made the decision to average over unexpected changes in seropositivity by age between 2014 and 2015 (see Appendix B.1.1). As Supplemental Figure B.1 shows, there are sometimes large discrepancies in the expected wave-like behavior of proportion seropositive as seen across years by ages. For example, the 2005 estimates show that 100% of the population over the age of 10 years old is seropositive, but the 2010 estimates show only 90% seropositivity among ages 15 to 35 years old. Based on the total proportion seropositive analysis by year (see Figure 2.4), the 2010-2015 data appears to better represent potential true changes based on known SIAs; this is likely due to larger sample sizes between 2010 and 2015. However, even between years 2014 and 2015, changes in the estimated proportion seropositive by age were not as wave-like as expected (e.g., a dip in immunity at age 5 years old in 2014 did not result in a dip of equal size or less at age 6 years old in 2015, etc.) (Supplemental Figure B.1). We avoided combining samples across time when a SIA took place, because SIAs caused sharp shifts of seropositivity among the targeted ages. The final model chosen to estimate the current age profile of immunity included an isotropic smoother on age with a cubic spline base and a logistic link function. The final model did not include a covariate for the year the sample was collected because we assumed that the limited sample represented the most current age profile of immunity. For more information on the model choice used to estimate current proportion seropositive by age, see Appendix B.1.2.
2.3.4 Indirect Empirical Estimates of Proportion Seropositive by Age - Birth Cohort Projection of Immunity

The ideal method for verifying the non-probability surveillance data would be to compare our age immunity profile estimates to age immunity profile estimates from population-based data that use random sampling for increased external validity. In the absence of population-based serum samples, we compared our direct empirical estimates of proportion immune to indirect empirical estimates of proportion immune, based on previously developed methods by Takahashi et al. (2015). The indirect method estimates the proportion immune for each birth cohort based on its experience of routine immunizations, SIAs, and natural infection. For example, if 80% of the birth cohort was routinely vaccinated and the cumulative measles attack rate was 75% among those unvaccinated, then 95% of the cohort would be estimated as immune (Takahashi et al., 2015).

We assumed that routine vaccination coverage rates were equivalent to the World Health Organization (WHO) and United National Children’s Fund (UNICEF) estimates for routine MCV1 coverage rates for Madagascar between years 1985 to 2014 (World Health Organization and UNICEF, 2015). SIA timing, age ranges, and vaccination coverage rates were extracted from WHO reported administrative estimates (World Health Organization, 2016b). Madagascar has administered four measles SIAs: 1) October 2004 which reported vaccination of 99% of children 9 months old (mo) through 14 years old (yo), 2) October 2007 which reported vaccination of 100% of children 9 mo through 4 yo, 3) October 2010 which reported vaccination of 93% of children 9 mo through 3 yo, and 4) October 2013 which reported vaccination of 92% of children 9 mo to 4 yo (World Health Organization, 2016b). Because all SIAs took place in the month of October, the SIA coverage rate contributed to theoretical estimates of the proportion immune for the year following the SIA. We assumed complete overlap between the probability of routine and SIA vaccination, so that where a cohort had experienced vaccination via an SIA as well as routine vaccination, coverage in that cohort was taken as be the highest of the two values. We assumed a maximum coverage rate of 95% in both SIAs and routine immunization, and applied a vaccina-
tion efficacy rate of 97% to both SIA and routine vaccination coverage rates (Boulianne et al., 1995).

We additionally estimated the probability of natural immunity for each birth cohort based on methods used by Takahashi et al. (2015). The probability of natural immunity as a function of age was estimated by assuming a constant hazard of infection for ages 1 to 100 years old. This is only an approximation, for younger ages the effect of natural immunity will be outweighed by the effect from vaccination. The base hazard rate for measles endemic years (prior to 1985) was set at -0.149, such that 95% of infections occur prior to 20 years of age (Takahashi et al., 2015):

\[ P(\text{infection by age } a) = 1 - \exp(-0.149 \times a) \]  \hspace{1cm} (2.1)

The base hazard rate in subsequent years, after the measles vaccine was introduced, was then scaled relative to the proportional decline in estimated measles incidence in each year relative to the mean incidence between 1981 and 1984, e.g., hazard rate in year \( t \) = 0.149 × (incidence in year \( t \) / mean incidence in 1981 to 1984). Estimated measles incidence was taken from the 2013 WHO measles burden estimates, calculated as per Simons et al. (2012). The probability of natural immunity for the first year of life was modified to account for maternal immunity, such that the mean of \( \exp(-0.45 \times a) \) individuals were considered immune, where \( a \) is age in months from 1 to 12 (Lessler et al., 2011b).

Both indirect and direct estimation techniques described above provide an estimate of the profile of immunity by age, however, indirect estimates require extrapolation from often poorly resolved parameters such as vaccination coverage; and further, unlike direct estimates, provide no clear measure of uncertainty. A summary of the differences, strengths, and limitations between these two estimation techniques can be found in Appendix B.1.3.
2.3.5 Simulate Measles Dynamics - Age-Structured Model

Building on our direct estimates of the current landscape of immunity over age, we deterministically simulated measles dynamics in Madagascar for 10 years to predict the size of a measles outbreak if an outbreak occurred. We used an established deterministic age-structured mathematical model, introduced in Metcalf et al. (2012a) and Metcalf et al. (2012b), to simulate measles transmission dynamics among the Malagasy population. An age-structured model controls for the oversampling of younger ages as described above. The key feature of the model was a matrix that at every discrete time-step defined transition from every possible epidemiological stage (e.g. maternally immune, susceptible, infected, recovered, and vaccinated) and age combination to every other possible epidemiological stage and age combination. The time-step was set to two weeks based on the generation time of measles, thus there were 24 discrete opportunities to move in and out of age and epidemiological stages per year. At every biweekly time period, the simulation outputs the number of individuals in each age and epidemiological stage.

We simulated nine different Malagasy populations based on three different starting populations, and three different levels of measles transmission. The starting populations were built from Madagascar’s population age structure and the estimated age immunity profile as described above. We extracted the 2015 total population size and population age structure from the United Nations’ estimates for Madagascar (United Nations, 2015). The population was stratified into 265 age groups (monthly up to 15 years old, yearly up to 100 years old) using smoothing splines. We then applied each estimated age immunity profile (mean and upper and lower bound of the 95% CI based on the penalized regression smoothing model as described above) to the population age structure in order to determine the number of susceptible and immune (i.e. recovered) individuals per age strata. The value of the assumed $R_0$, a summary measure of transmission, combined with the population size $N$, sets the intensity of transmission with $R_0 = \bar{\beta}N$, where $\bar{\beta}$ is average transmission. $R_0$ is the basic reproductive number and is qualitatively defined as the average number of people a typical infected individual will infect in a fully susceptible population. Estimates of $R_0$ for measles range
globally from 5 to 20 (Anderson and May, 1991); we assumed three different $R_0$ values (10, 15, and 20).

Additional data inputs required by the age-structured models included age contact rates, the predicted routine measles vaccination coverage, and Madagascar’s birth and survival rates. We assumed a pattern of contact over age proportional to what was measured in Great Britain based on the POLYMOD diary study in Europe (Mossong et al., 2008). The predicted level of routine vaccination coverage for measles containing virus dose one was 64%. We value was extracted from the WHO/UNICEF 2014 estimate, and was kept constant over time (World Health Organization and UNICEF, 2015). Population survival rates per age class were estimated by matching the region’s 2013 estimated life expectancy at birth (65 years old; The World Bank Group, 2015), to the United Nations ‘Coale-Demeny East’ life table model (UN Population Division, 2010). We assumed constant rates of mortality by age over time. Crude birth rates between 2015 and 2025 were extracted from the United Nations Population Divisions probabilistic population projections for Madagascar (United Nations and Department of Economic and Social Affairs and Population Division, 2015). It is important that we included realistic projections of fertility in the model because the local dynamics of measles elimination are affected more by the absolute number of susceptible individuals than the proportion of the population that is susceptible (Ferrari et al., 2013).

The model assumed spatially homogenous mixing among the Malagasy population. The model also assumed seasonal forcing, where the intensity of transmission varied over each year (affecting all age groups the same), by setting $\beta_t = \bar{\beta}(1 - \alpha \cos(2\pi t/24))$ for each time step $t$; and $\alpha$ was set to 0.35 which reflects the factor by which the peak and lowest value of $R_0$ differ from the average. We introduced an infected individual of age 10 years old (the age with the highest assumed contact rates) into the simulated populations at bi-week 2 in order to instigate a measles outbreak. Additional model methods are presented in detail in Metcalf et al. (2012a) and Metcalf et al. (2012b), and are therefore not further discussed here.
2.3.6 Explore the Effect of Measles SIA Scenarios

SIAs that target age groups of high susceptibility can be effective tools to reduce the risk of a measles outbreak and move towards elimination (Babad et al., 1995; Osborne et al., 1997). We ‘implemented’ a range of SIA scenarios to the current directly estimated age profile of immunity (mean and 95% CI), and simulated measles dynamics forward using the age-structured model described above.

Sixteen different SIA scenarios were assessed. Each scenario differed by four targeted age group and four levels of vaccination coverage. The targeted age groups included two classically targeted age groups (9 months old to 5 years old, and 9 months old to 15 years old), and two non-classically targeted age groups (9 months old to 10 years old, and 9 months old to 20 years old). The four levels of vaccination coverage were 70%, 80%, 90%, and 95%, which ranged from conservative estimates (70% and 80%) to past reported SIA coverage estimates, which are at the higher end (90% and 95%) (World Health Organization and UNICEF, 2015). In our model, the SIA was implemented at bi-week 1 and the infected individual was introduced at bi-week 2. As in the indirect estimates of proportion immune above, we assumed: i) complete overlap between the probability of routine and SIA vaccination, ii) maximum coverage of 95% in both SIAs and routine immunization, and iii) a vaccination efficacy rate of 97% to both SIA and routine vaccination coverage rates (Boulianne et al., 1995). We also accounted for the interference of maternal antibodies on vaccination success among children under the age of 1 years old, by assuming that the probability of vaccination success follows a logistic function $1/(1 + \exp(-3.77 + 0.59a))$, where $a$ is age in months 1 to 12 (Lessler et al., 2011b). We treated Madagascar as one large population without spatial demarcations; therefore, vaccination rates were applied homogeneously across the Malagasy population.

To explore the potential of each measles SIA to prevent a measles outbreak and move towards measles elimination, we: i) estimated population immunity levels post SIA and compared it to the
the critical immunity threshold, $p_c$, ii) estimated the effective reproductive number, $R_{eff}$, post SIA, and compared it to elimination threshold of one, and iii) estimated outbreak size post-SIA and compared to outbreak size if no SIA was implemented.

First, we compared the critical immunity threshold, $p_c$, to the estimated population immunity for each SIA scenario. Estimates were based on the population at bi-week 1, i.e., after the SIA was implemented but before the infected individual was introduced into the population. Population immunity for each SIA was estimated using an age-weighted average of the proportion seropositive for different starting age profiles of immunity (mean and 95% CI based on direct estimates). The critical immunity threshold is defined as:

$$p_c = 1 - \left( \frac{1}{R_0} \right)$$  \hspace{1cm} (2.2)

We assumed three different values of $R_0$: 10, 15, and 20, such that $p_c$ was estimated as 90%, 93.3%, and 95%. In the simplest analysis, $p_c$, is the level of population immunity required to achieve herd immunity (Fine, 1993); if population immunity is maintained below the critical immunity threshold over time, then the virus has the potential for elimination (Anderson and May, 1985; Fine, 1993; Kermack and McKendrick, 1927). While $p_c$, as defined above, is used as a benchmark for informing policies aimed at elimination (Gay, 2004), in reality, the exact threshold for elimination will be sensitive to the details of the pattern of transmission over age in combination with the age structure (Gay et al., 1995; Anderson and May, 1985). For example, achieving levels of immunity above $p_c$, as defined above, in younger age groups where transmission is high, can allow for lower levels of immunity in older age groups and potentially a higher value of $p_c$ overall (Fine, 1993).

The effective reproductive number, $R_{eff}$, is a related measure used for assessing elimination. $R_{eff}$ is the average number of secondary cases per typical infected individual. Our age-structured model used next generation methods to define $R_{eff}$ as the dominant eigenvalue of the who acquires infection from whom matrix (pattern of transmission over age) multiplied by the proportion susceptible in each age group. The $R_{eff}$ must remain below one in order to eliminate
measles. Therefore, we additionally estimated $R_{eff}$ over a simulation time period of 10 years to assess the potential of SIAs to reduce population susceptibility and move towards elimination. We directly compared estimates of $R_{eff}$ immediately following the SIA (bi-week 1) to the proportion of the population susceptible immediately following the SIA (bi-week 1) to highlight differences between the two elimination thresholds (proportion susceptible $< (1 - p_c)$ and $R_{eff} < 1$).

Lastly, we estimated the measles outbreak size assuming an infected individual was introduced into the simulated population post SIA (bi-week 2), and compared it to the outbreak size if no measles SIA was implemented. Given the difficulty in defining the parameters of an ‘outbreak’, specifically for large-coverage SIAs, we simply compared the the number of measles cases that occurred within three years of the introduction of an infected individual (bi-weeks 2 - 96).

### 2.4 Results

Figure 2.3 displays the number of measles and rubella cases identified via IgM antibodies diagnostic testing between November 2004 and December 2015 in Madagascar. Fifty-two of the 4446 serum sample were positive for measles IgM antibody. All measles cases in the dataset were serotyped; therefore no cases were epidemiologically linked. This is an example of dwindling measles incidence data, when there is little more information to gather from measles IgM serological data because cases are rare (Figure 2.3). The 52 measles case samples were removed from the dataset for the remainder of the analysis to minimize oversampling of measles seropositive individuals from natural immunity. The total sample size that remained was 4394 samples.

Figure 2.4A shows the proportion of the population under the age of five years old that was immune to measles between years 2005 and 2015 for both the direct and indirect empirical estimates. The indirect estimate of proportion immune under the age of five years old between years 2005 and 2015 ranged from 73.1 to 91.3%; the direct mean estimates were smaller, ranging from 61.8 to 76.4%. If
we assume that the empirical estimates are correct, these differences may be due to over-reporting of vaccination coverage. If we assume that the indirect estimates are correct, these differences may be due to biased sampling of susceptible individuals. The direction of change of the proportion seropositive growth rate between years 2005 and 2009 from the direct estimates did not match the direction of change of the proportion seropositive growth rate between years 2010 and 2015 for the indirect estimates. We expected to detect an increase in the proportion immune after the SIA that took place in October of 2007, which targeted children from 9 months to 5 years old. Our inability to detect expected changes in immunity between 2005 and 2009 was likely due to small sample size and potentially due to poor representation of the general population under five years old within this time period. In contrast, the direction of change of proportion seropositive between 2010 and 2015 from the direct estimates did match the direction of change of the proportion seropositive growth rate between 2010 and 2015 for the indirect estimates. The direct estimates captured the footprint of the SIAs that took place among children less than five years old in October 2010 and in October 2013, with an increase in the proportion seropositive in 2011 and 2014, and decreases in alternative years. These expected shifts in the empirical data were likely the result of better
Figure 2.4: Estimated proportion seropositive 2005-2015. a) The proportion seropositive of children under 5 years old between 2010 and 2015. b) The proportion seropositive of the total population between 2010 and 2015. The blue dots and line represent the indirect estimates based on the birth cohort projection of immunity. The black dots and line represent the mean direct estimates based on the nested serological survey and the grey shaded area represents the 95% confidence intervals of the direct estimates based on serological data (N=4083).

representations of the data due to larger sample sizes for years 2010 and 2015.

Figure 2.4B displays the proportion of the total population of Madagascar who are immune to measles between years 2005 and 2015. Similar to the under five estimates, the direction of change observed in the proportion seropositive growth rates between 2010 and 2015 from the direct estimates matched the direction of change in the proportion seropositive growth rates for years 2010 and 2015 for the indirect estimates, however this observation did not hold true for years 2005 and 2009. The indirect estimates generally were more stable and ranged by only 2%, while the direct mean estimates ranged by 14% between 2005 and 2015. The correlation between indirect and direct population immunity estimates was 0.28. The indirect estimates fell within the 95% confidence bounds of our direct estimates with the exception of 2013, highlighting some congruence.
Figure 2.5 displays the estimated current proportion seropositive by age. We show the observed serological survey data grouped by age in months, the direct estimates (based on model predictions from 1083 serological samples), and the indirect estimates (based on birth cohort projection of immunity). The nested serological data, represented by the grey dots, shows that by age 40, the whole sample population was estimated to be seropositive or immune to measles. Prior to age 40, there was a general increase in the proportion seropositive by age with an exception of age 12 and 28 years, when there was a general drop in immunity. The fitted line, based on a model with a penalized regression smoother on age captured this pattern well (see Appendix B.1.2 information on model choice). The solid dark blue line shows the mean estimated proportion seropositive and the dashed dark blue lines show the 95% confidence intervals. Over the age of 37 years, the lower 95% confidence bound dropped to zero because there was very little information for older ages. We therefore modified the lower 95% confidence bound to increase to 0.999 proportion seropositive, as represented by the dashed pink line.

The age profiles of immunity differed between direct and indirect estimates (Figure 2.5). The indirect estimates predicted high immunity of 95% from age 8 to age 26 as a result of a 95% vaccination coverage rate in the 2004 catch-up campaign that extended to age 15 years, and in the 2007 SIA that extended to age 5 years. The indirect estimate also predicted a large pocket of susceptible individuals just outside the 2004 SIA age range (individuals who would be between ages 27 and 32 years old in 2015), and were not eligible for any SIAs, and who experienced relatively low routine coverage. In contrast, the direct estimates suggested that the 2004 SIA was heterogeneously applied across ages, an important and yet rarely described features of this type of intervention. Peak coverage appears to have been achieved for children whose age was around the middle of the target age range, with lower seropositivity seen in both 12 and 28 year olds.

Based on serological data, we estimated that 84.9% of Madagascar’s current population is immune to measles (95% CI: 75.3%, 89.8%). For a population of approximately 25 million individuals (United Nations 2015), this means that 3.75 million Malagasy remain susceptible to measles.
Figure 2.5: Estimated current proportion seropositive by age. The grey dots represent the observed seroprevalence per age in months based on the nested serological data. The dark blue solid and dashed lines represent the mean and 95% CI of the estimated proportion seropositive based on the serological data (N=1083). The light blue line represents the indirect estimates of proportion seropositive by age.

(95%CI 6.25 to 2.5 million). The indirect estimate of the proportion immune (88.9%) fell just inside the 95% confidence bounds, and differed from the mean direct estimate by 4%. Both the direct and indirect estimates showed that population immunity in Madagascar is below the simple unstructured estimate of the critical immunity threshold for herd immunity (90-95%) ([Anderson and May 1985] [Moss and Griffin 2006]). Therefore, Madagascar is at risk of experiencing a significant measles outbreak.

Assuming homogenous spatial mixing across Madagascar, we estimated outbreak sizes ranging from 738,000 to 5.4 million. Outbreak sizes was dependent on the assumed value of $R_0$ and the assumed age profile of immunity (see Appendix B.2.1). However, various lines of evidence suggest that the assumption of homogenous spatial mixing is inappropriate, e.g., the presence of remote
districts suggested by analysis of rubella incidence in Madagascar (Wesolowski et al., 2016), and the fact that measles cases have occurred in Madagascar without any major outbreaks. Therefore, an observed measles outbreak would be much smaller than predicted here, because the virus would spread less quickly or evenly in a real world scenario. One approach to capturing the reduction in incidence as a result of heterogeneous spatial spread is to use a mapping between susceptibility and incidence based on Kalman Filtering based estimates of measles burden (Simons et al., 2012). However, given that our focus is specifically on the nuance of the impact of age targeted SIAs, this simplification would obscure much of the variation of incidence. Therefore, we retained the homogenous framing and used order of magnitude to compare outbreak sizes between this no SIA scenario and SIA scenarios (see the last paragraph of the results section for this analysis).

Figure 2.6 displays the total proportion seropositive per SIA strategy by applying different SIA scenarios to the serological data direct estimates of the current immunity profile over age (mean and 95% CI) in Madagascar. We present the results by comparing the mean and 95% confidence intervals of total population immunity per SIA strategy to the critical immunity threshold value when \( R_0 \) was 10, 15, or 20. Despite the fact that up to 30% of cases occurred among children up to 5 years old if no SIA campaign was implemented (results not shown), an SIA campaign that only targeted this age group would likely not prevent a measles outbreak. If the ‘true’ proportion of immune individuals was close to our mean estimate, 85.3%, a SIA campaign targeting children up to age 15 years with at least 90% coverage prevented a measles outbreak if \( R_0 \) was 10 or 15. If \( R_0 \) was 20, there was a measles outbreak; however, the magnitude of the outbreaks with these SIAs in place were much smaller than if no SIA took place (see the last paragraph of the results section for this analysis).

Figure 2.7 displays the estimated \( R_{eff} \) by proportion susceptible for all SIA scenarios by assumed value of \( R_0 \) across columns. The filled in dots represent the no SIA scenario, the other shapes represent the 16 different SIA scenarios. The colors represent different starting age profiles directly estimated from serological data; black is the mean estimate, red is the modified 2.5% confidence
Figure 2.6: The total proportion immune (seropositive) post SIA campaign by SIA strategy. The black dots represent the mean proportion seropositive and the horizontal black lines the proportion seropositive 95% confidence interval. The vertical lines represent the critical immunity threshold, \( p_c \), associated with the \( R_0 \) values of 10, 15, and 20. The rows represent different SIA scenarios where age targets range between 9 months old (mo) and 5, 10, 15, and 15 years old (yo), and vaccination coverage rates range between 70, 80, 90, and 95%.

lower bound (lb), and blue represents the 97.5% confidence upper bound (ub). Each cluster of similar shapes and colors per graph represents vaccination coverages of 70%, 80%, 90%, and 95% (in descending order from right to left) per SIA age group target and starting age immunity profile. The grey dashed horizontal line represents the elimination threshold of \( R_{eff} \) (i.e., 1), and the vertical dashed line represents the elimination threshold of proportion susceptible (i.e., \( (1-p_c) \)).

It is important to reiterate that \( R_{eff} < 1 \) represents a more exact elimination threshold than the population proportion susceptible \( < (1-p_c) \) in our analysis, because it takes into account the age contact matrix in combination with the population age structure. Points that fell in the bottom left and top right quadrant of each graph in Figure 2.7 represent agreement of the elimination thresholds. Points that fell in the bottom right quadrant imply that \( p_c \) may be a conservative estimate for these scenarios, and that the exact elimination threshold for the population proportion susceptible may be greater than \( (1-p_c) \) (Figure 2.7). Points that fell
in the top left quadrant should receive the most focus (Figure 2.7); these were scenarios where we estimated $R_{eff}$ was greater than 1, even though the population proportion susceptible was less than $(1 - p_c)$. For example, assuming $R_0$ was 10 and that the ‘true’ age immunity profile was equal to the directly estimated 97.5% upper confidence bound of the age immunity profile, SIA\s that target up to 5 years old and up to 10 years old would not be sufficient to reduce $R_{eff}$ below 1 and prevent a measles outbreak across all levels of vaccination coverage, contrary to Figure 2.6. The same was true for the following scenarios: i) assuming $R_0$ was 10, scenarios where the ‘true’ age immunity profile was equal to the directly estimated mean age immunity profile and SIA\s that targeted children up to 10 years old with vaccination coverage of 90 or 95%, ii) assuming $R_0$ was 15, scenarios where the ‘true’ age immunity profile was equal to the directly estimated 97.5% upper confidence bound of the age immunity profile, for SIA\s that target children up to 10 years old with 80, 90, or 95% vaccination coverage. Figure 2.7 also shows clearly that if no SIA was implemented (filled in dots in each graph), the population was at risk of a significant measles outbreak, and not en route towards elimination based on the $R_{eff} < 1$ threshold.
We additionally estimated $R_{\text{eff}}$ for the entire 10 year simulation (see Appendix B.2.2). These figures indicated that in a growing population such as Madagascar, even if an SIA successfully reduced $R_{\text{eff}}$ below 1 immediately post-SIA, the influx of susceptible from births who miss the routine measles vaccine continued to increase $R_{\text{eff}}$, and the population was again at risk of a measles outbreak. Our findings highlight the need for vaccination programs to remain vigilant as we move toward measles elimination.

The expected size and duration of an outbreak depended directly on $R_{\text{eff}}$, such that larger values of $R_{\text{eff}}$ resulted in larger and longer outbreaks (de Serres et al., 2000). Figure 2.7 clearly shows that measles outbreaks were significantly larger and longer if no SIA was implemented (see Appendix B.2.1 for time series of measles cases per SIA scenario). Table 2.1 reveals the percent reduction in the number of measles cases over three years for each SIA scenario as compared to the no SIA scenario for each value of $R_0$ and estimated age immunity profile (mean and 95% CI). We estimated that a SIA that successfully vaccinated 95% of children between the age 9 month old to 5 years old reduced the number of measles cases in a measles outbreak by 17 to 45%, depending on $R_0$ and the true age immunity profile. We also found that a SIA that successfully vaccinated 95% of children between the age 9 months old to 15 years old reduced the number of measles cases in a measles outbreak by 65 to 100%, depending on $R_0$ and the true starting age immunity profile.

### 2.5 Discussion

Fifty-two of the 4446 serum samples tested positive for measles IgM antibodies. Our data most likely under-reported the true number of measles cases that occurred because not all patients with measles presented for care, and of those who do present for care, not all have serum collected and sent for testing (CDC, 2014). Despite the fact that measles is much more transmissible than rubella ($R_0$ as much as 5 times as high, (Anderson and May, 1991)), 1326 rubella cases were reported, and rubella outbreaks reached the size of 43 individuals in a single month. This suggests that i) the
Table 2.1: Percent reduction in the number of measles cases over three years after introduction of an infected individual for each SIA scenario compared to if no SIA was ‘implemented.’ Estimates are based on simulations from the age-structured measles dynamics model [Metcalf et al. 2012b,a], for three different assumed values of $R_0$, and three different starting age immunity profiles per our direct empirical estimates (mean, modified 2.5% confidence lower bound (lb) and 97.5% confidence upper bound (ub)). The rows represent different 16 SIA scenarios where age targets range between 9 months old (mo) and 5, 10, 15, and 20 years old (yo), and vaccination coverage rates range between 70, 80, 90, and 95%.

<table>
<thead>
<tr>
<th>SIA option</th>
<th>$R_0=10$</th>
<th></th>
<th></th>
<th>$R_0=15$</th>
<th></th>
<th></th>
<th>$R_0=20$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lower bound</td>
<td>mean</td>
<td>upper bound</td>
<td>lower bound</td>
<td>mean</td>
<td>upper bound</td>
<td>lower bound</td>
<td>mean</td>
</tr>
<tr>
<td>no SIA (ref)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9mo-5yo, 95%</td>
<td>19.7425</td>
<td>26.1160</td>
<td>45.4039</td>
<td>18.2343</td>
<td>23.0612</td>
<td>27.9863</td>
<td>17.0821</td>
<td>22.6715</td>
</tr>
<tr>
<td>9mo-10yo, 70%</td>
<td>36.8472</td>
<td>54.6240</td>
<td>96.5899</td>
<td>29.8608</td>
<td>40.9112</td>
<td>54.0509</td>
<td>26.2172</td>
<td>36.9131</td>
</tr>
<tr>
<td>9mo-10yo, 80%</td>
<td>41.3109</td>
<td>60.6726</td>
<td>98.6278</td>
<td>33.5974</td>
<td>45.8720</td>
<td>60.0345</td>
<td>29.9870</td>
<td>41.4320</td>
</tr>
<tr>
<td>9mo-10yo, 90%</td>
<td>45.4232</td>
<td>66.1851</td>
<td>99.4033</td>
<td>37.4966</td>
<td>50.5633</td>
<td>65.3668</td>
<td>33.6407</td>
<td>45.8643</td>
</tr>
<tr>
<td>9mo-10yo, 95%</td>
<td>47.9638</td>
<td>68.6523</td>
<td>99.6084</td>
<td>39.4309</td>
<td>52.6399</td>
<td>67.7823</td>
<td>35.3284</td>
<td>48.0796</td>
</tr>
<tr>
<td>9mo-15yo, 70%</td>
<td>80.9617</td>
<td>99.9906</td>
<td>99.9996</td>
<td>55.2867</td>
<td>85.1452</td>
<td>99.9889</td>
<td>45.4577</td>
<td>88.8625</td>
</tr>
<tr>
<td>9mo-20yo, 70%</td>
<td>99.0534</td>
<td>99.9995</td>
<td>99.9997</td>
<td>69.4894</td>
<td>98.8676</td>
<td>99.9991</td>
<td>57.0745</td>
<td>78.8507</td>
</tr>
<tr>
<td>9mo-20yo, 80%</td>
<td>99.9946</td>
<td>99.9998</td>
<td>99.9998</td>
<td>82.5364</td>
<td>99.9984</td>
<td>99.9998</td>
<td>67.7267</td>
<td>98.0618</td>
</tr>
</tbody>
</table>

surveillance system is functioning, and ii) that the rarity of measles cases is not purely an issue of under-reporting.

As a result, we conclude that the measles vaccination program in Madagascar has been successful to date. Since the introduction of the vaccine, measles cases have decreased from endemic levels, and Madagascar has avoided large reported measles outbreaks post vaccine introduction that have been seen in other countries due to the accumulation of susceptible individuals [Chen et al. 1994; Cutts et al. 2013; Kidd et al. 2012; Minetti et al. 2013b,a; Shibeshi et al. 2014]. While the measles control program has worked to successfully decrease measles cases, this very success has resulted in a declining power of the measles and rubella surveillance program to reveal measles disease dynamics. In the absence of sufficient measles incidence data, public health officials must rely solely on estimated vaccination coverage to assess the risk of measles outbreaks, and to guide SIA timing and target age ranges [Takahashi et al. 2015]. This analysis proposes filling the information gap by leveraging existing samples from the measles and rubella surveillance system to conduct a measles IgG serological survey. Measles IgG serological surveys can be used to estimate...
the population immunity profile of measles by age in the wake of dwindling measles incidence data (Gay et al., 1995; Hens et al., 2012; Vyse et al., 2006).

The ideal method to verify that nested serological survey data produces generalizable estimates of the age profile of immunity would be to compare this data to age immunity profile estimates from population-based or community-based data. Unfortunately, there is no population- or community-based serological data publicly available in Madagascar to date. It is therefore important to discuss how our nested serological sample may differ from a population-based sample. If we assume that non-measles related fever and rash symptoms are random across individuals via measles seropositivity status, then the sample may overestimate population immunity because the utilization of health services is typically heterogeneous across populations. For example, geographic accessibility to health facilities also increases the utilization of health care services such as vaccination or treatments for fever and rash symptoms (Buor, 2003; Gething et al., 2004; Tanser et al., 2006). As a result, our data may oversample individuals who are measles seropositive due to vaccination.

The sample may also overestimate measles population immunity if non-measles related fever and rash symptoms are not random across the population via measles seropositivity status. The measles virus can suppress the immune system for several weeks after infection, resulting in increased susceptibility to secondary infections (Mina et al., 2015; Griffin, 2010). The most commonly documented secondary infections to measles are those that cause diarrheal diseases and pneumonia (Nelson and Williams, 2014), but measles infection can also place individuals at risk of rubella, chickenpox, or other fever and rash causing pathogens. If these pathogens have annual outbreaks it could, in theory, cause age bias oversampling of measles seroprevalence, because these pathogens tend to have characteristic ages of infection. However, outbreaks of other fever and rash causing pathogens among measles infected individuals are unlikely to play any large role in measles immunity bias in the sample, given that the data suggests very low incidence of measles since 2004.
Given the lack of population-based serological data, we used other sources of available data, and compared our nested serological data direct estimates of age immunity profile to indirect estimates of age immunity profile from the birth cohort projection of immunity (Takahashi et al., 2015). Neither our mean direct or indirect estimates can be validated as ‘true’; however each estimated the total proportion immune to within five percentage points of each other, even though the age distributions were different. More importantly, both the indirect and direct estimates of proportion seropositive were less than $p_c$, and direct estimates of $R_{eff}$ were less than one, suggesting that Madagascar is at risk of a major measles outbreak (Anderson and May, 1985; Fine, 1993; Kermack and McKendrick, 1927; Moss and Griffin, 2006).

Nested serological surveys have considerable potential as an alternative to expensive and time-consuming population-based serosurveys. Additionally, compared to indirect estimates based on birth cohort projection of immunity analysis, they allow probabilistic estimates of measles seroprevalence by age with standard errors. Understanding age profile of immunity is an invaluable tool for identifying age gaps in immunity that inform SIA target age groups (Gay et al., 1997; Nigatu et al., 2008). This resource is especially important to avoid late age outbreaks that have recently taken place in sub-Saharan Africa due to an unknown build up of susceptible individuals in adulthood (Cutts et al., 2013; Kidd et al., 2012; Minetti et al., 2013b,a; Shibeshi et al., 2014).

Our analysis points to the usefulness of nested IgG serological surveys within the measles-rubella surveillance system in Madagascar. With this data we were able to estimate age immunity profiles, and predict the risk of outbreaks and outbreak size. Direct estimates revealed that 15%, ranging from 10 and 25%, of Madagascar’s population is susceptible to measles. Other than expected susceptibility among infants less than one year old (Mckee et al., 2015), we estimated dips in immunity around 12 years old and again around 28 years old. We estimated measles outbreaks sizes assuming no SIA, and also based on a range of SIA scenarios; the resultant outbreak sizes are shown in Appendix B.2.1. Given the likely unrealistic assumption of spatial homogenous mixing in Madagascar, we focused on order of magnitude to compare outbreak sizes between the no SIA
scenario and SIA scenarios.

Simulating the effects of different SIA scenarios on starting age profiles of immunity, we found that SIAs that targeted children between the ages of 9 months old and 5 years old were not sufficient to effectively reduce the number of susceptible individuals below the 5-10% threshold, or $R_{eff}$ less than one (Anderson and May 1991; Bartlett 1956; Fine 1993; Kermack and McKendrick 1927; Moss and Griffin 2006); however, we found that they did serve to reduce the outbreak size by 17 to 47%. These estimates were based on the model assumptions that the age contract matrix in Madagascar is proportional to that found in Great Britain via the POLYMOD diary study (Mossong et al. 2008). The general findings in the POLYMOD study of high rates of contact among assortative ages and between parent and child ages have also been replicated in China and Vietnam (Read et al. 2014; Horby et al. 2011). However if the age contact rates in Madagascar are unique, it would change the magnitude of effect on SIAs based on the targeted age range.

Our analysis showed that SIAs which target children between ages 9 months and 15 years old were most efficient in reducing susceptibility per additional age added to the SIA target. As such, this SIA is predicted to be successful in moving the population towards measles elimination by significantly reducing the number of susceptible individuals. If $R_0$ was 15 or less, this SIA prevented a major measles outbreak in Madagascar, assuming homogenous vaccination coverage, and in the very least significantly reduced the number of measles cases between 80 and 100%. If $R_0$ was 20, an SIA that targeted individuals up to 15 years old reduced the number of measles cases in an outbreak by 65 to 100%. Fortunately, there is an available opportunity to conduct a measles SIA targeting children up to age 15. Madagascar has announced a plan to introduce the rubella-containing vaccine via the Measles-Rubella (MR) vaccine; this introduction is predicted to be successful at effectively reducing the country’s Congenital Rubella Syndrome (CRS) burden. The MR vaccine introduction is eligible for GAVI (i.e., Global Alliance for Vaccines and Immunizations) financial support, which will fund an initial rubella catch-up campaign by providing the measles-rubella vaccine for all individuals ages 9 months to 15 years old (Global Alliance for Vaccines and Immunizations 2014).
In conclusion, as countries proceed with measles elimination programs, measles outbreaks will remain a potential threat over time due to the build-up of susceptible individuals from missed vaccines and births (Wallinga et al., 2005). The size of an outbreak will primarily depend on the number of susceptible individuals (Bartlett, 1956). Measles control programs must remain vigilant in order to reduce the risk of outbreaks by determining population immunity age profiles. Measles IgG serological data via surveillance serum samples is a potentially cost-effective data source that can be used to assess the threat of outbreaks. Further investigation comparing surveillance data to population-based data is recommended before generalized investment and use of this data source.
3.1 Abstract

Introduction: Rubella is predominantly a mild childhood infectious disease, but infection during early pregnancy may cause fetal death, spontaneous abortion, or the birth of an infant with congenital rubella syndrome (CRS). Rubella and CRS are highly under-reported via case-based national surveillance systems. We rely on serological surveys to estimate rubella epidemiological parameters including the incidence of CRS, and to determine the effect of rubella vaccine introduction. However, it is unknown how typical serological survey data, which often have limited scopes of age ranges and sample sizes, may affect estimation of these key parameters. Methods: This analysis used simulation tools to explore the strength of serological data in estimating epidemiological parameters, and importantly to characterize biases associated with analytic methods and serological survey designs, as well as overall levels of uncertainty. To do this, we simulated three populations representing a range of rubella transmission levels, and extracted variables of interest: the age immunity profile, the age specific force of infection, rubella’s basic reproductive number, and CRS incidence rate. We then systematically explored biases from i) analytic methods, ii) chosen survey age-groups, and iii) sampling bias, by comparing ‘true’ parameter values (extracted from simulated
populations) to parameter estimates. **Results:** Broadly, our results indicated that i) even for pre-vaccination stable rubella dynamics, flexible semiparametric approaches to analyzing serological data better capture true estimates than non-parametric approaches, ii) while the two-age-group piecewise constant method used by the WHO to estimate global CRS burden certainly biases CRS incidence estimates, the magnitude and direction of bias is highly variable across different sampling survey designs, iii) parameter estimates from the frequently used serological design of five year age-groups can be considerably improved by including one or two additional age breaks at 13 years old and 25 years old, as this reduces the magnitude of parameter estimate bias, and iv) a larger sample size is needed to increase precision of parameter estimates if $R_0$ of rubella is suspected to be greater than five. **Conclusion:** As new funding sources become available for the Measles and Rubella containing vaccine (e.g., funds from the Global Alliance of Vaccines and Immunizations), serological surveys will become increasingly important tools for evaluating the burden of CRS and the impact of vaccination campaigns. Understanding the inherent biases in parameter estimates from serological data in complex infectious disease dynamics is an important endeavor for researchers and public health officials who are making policy recommendations based on analyses of serological data. Here we characterize some of these biases.

### 3.2 Introduction

Rubella typically presents as a mild febrile rash illness in children; however, rubella infection in pregnant women can cause detrimental outcomes such as spontaneous abortion, fetal death, and the birth of an infant with birth defects (i.e., congenital rubella syndrome (CRS)) (Greenberg et al., 1957; Enders et al., 1988; Miller et al., 1982). Rubella-containing vaccine (RCV) is a safe and effective vaccine (Reef and Plotkin, 2013) that with high uptake can be used to successfully eliminate endemic rubella and CRS cases, as has been demonstrated in the WHO region of the Americas (Figueroa et al., 2014). As of 2015, RCV had been introduced into the national immunization schedules in 149 of 194 WHO member states (World Health Organization, 2016c). For the remaining countries considering introducing the rubella vaccine, the first step is understanding the national burden of CRS, and the second step is examining the potential of RCV introduction to reduce this burden. There are no reliable direct estimates of CRS incidence for many target regions due to poor country-wide surveillance infrastructures, and due to variation in laboratory techniques and medical definitions used to identify CRS (Lawn et al., 2000). We therefore rely upon serological data and mathematical models to estimate CRS incidence (Vynnycky et al., 2016), and explore the effect of RCV introduction (Wesolowski et al., 2016).
Age structured Immunoglobulin G (IgG) serological surveys provide a window into the landscape of population immunity for immunizing infections such as rubella. Rubella IgG antibodies are a marker of prior infection or immunization, and an indicator of immunity. In countries that have yet to introduce RCV, age structured IgG serological surveys allow for estimation of unknown transmission dynamic parameters such as the age-specific force of infection (ASFOI) (i.e., the rate at which individuals susceptible to rubella will become infected with rubella) (Chakravarty et al., 1976; Ximenes et al., 2014). Assuming known population demography, we can estimate the burden of CRS from the ASFOI and immunity profile of reproductive ages (Vynnycky et al., 2016).

Serological surveys can also be used to explore the consequences of RCV introduction. Empirical and theoretical research have shown that levels of RCV coverage below a critical threshold can result in an increase in CRS incidence in the short term, by increasing the age of infection without sufficiently reducing the incidence of rubella cases (Anderson and May, 1983; Knox, 1980; Papanicolaou et al., 1999). The basic reproductive number ($R_0$) is a summary measure estimated from the ASFOI (Farrington et al., 2001), and is necessary for determining the critical threshold of vaccination coverage level required to prevent short and long-term increases in CRS post RCV introduction (Anderson and May, 1991; Lessler and Metcalf, 2013).

An ideal cross-sectional serological survey would have highly resolved age classes, a clear knowledge of the functional form of age immunity, and a very large sample size in order to accurately and precisely estimate parameters of interest. In reality, both underlying information and this scope of data is rare due to limited resources. For example, Vynnycky et al. (2016) published a review of serological data and used mathematical models to estimate that 105,000 global incident cases of CRS occur annually. These estimates are based on a two-age-group (13,50] piecewise constant analytic approach, likely chosen due to non-consistent serological survey designs, some with limited scope of age ranges. Here, we aimed to assess serological surveys bias in the inference of key epidemiological parameters: rubella seroprevalence by age, rubella ASFOI, rubella $R_0$, and CRS incidence rate. We organized our analysis around three types of bias that can affect our estimation: bias from the analytic method, bias from survey defined age-groups, and bias from sampling (Table 3.1). We also sought to identify the simplest survey design requiring the fewest age-groups.
resulting in minimal parameter estimation bias in order to inform the design of future rubella serological surveys in areas where the vaccine has yet to be introduced.

Table 3.1: Types of parameter estimation biases from serological survey data.

<table>
<thead>
<tr>
<th>Source of Bias</th>
<th>Description of Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytic Method</td>
<td>Potential bias in accuracy of parameter estimates emerges from model mis-specification, i.e. the chosen functional form is insufficiently and cannot capture the true age pattern of seroprevalence or force of infection, even using perfect data at every age in years.</td>
</tr>
<tr>
<td>Survey Defined Age Group</td>
<td>Inherent bias in precision of parameter estimates emerges from: 1) misrepresentation of underlying functional forms (e.g. breaking apart sharp changes in transmission vs. grouping them), 2) oversampling of younger ages within each group in a growing population.</td>
</tr>
<tr>
<td>Sampling</td>
<td>Inherent bias in accuracy and precision of estimates comes from binomial sampling process, designated sample sizes, and parameters’ distribution across age.</td>
</tr>
</tbody>
</table>

3.3 Methods

To meet our objectives, we compared ‘true’ epidemiological parameters based on three simulated populations to estimated epidemiological parameters from different analytic methods and survey designs. The epidemiological parameters of interest were seroprevalence by age, the ASFOI, the basic reproductive number ($R_0$) of rubella, and the CRS incidence rate. We explored three different analytic methods for analyzing serological data (penalized regression splines, fractional polynomials, and two-age-group piecewise constants), 21 different survey design defined age-grouping schemes, and three different survey design sample sizes. The accuracy and precision of estimates were evaluated in a stepwise approach in order to tease apart the effect of each bias. First, we assessed the role that analytic methods play in potentially biasing estimates of epidemiological parameters by using perfect seroprevalence data at every age in years so that the only source of bias was due to mismatches resulting from the functional form specified by the analytic method. Second, we used perfect seroprevalence data for each age-group, i.e., the average proportion seropositive per defined age-group assigned to the mid-age, to explore the effect of biases from both
analytic methods and grouping ages as defined by a serological survey design. Last, we examined biases from analytic methods, survey defined age-grouping, and sampling across different sample sizes, by estimating epidemiological parameters using ‘realistic’ data from simulated serological surveys, where a finite number of samples were sampled from each age-group. In each analysis, we compared our estimated parameters to the true parameter values (known because we simulated the population’s rubella dynamics). We also compared parameter estimates to the best estimates per analytic method to disentangle the various sources of bias. See Table 3.2 for a layout of the stepwise approach to exploring these types of biases.

Table 3.2: Stepwise approach to exploring three types of parameter estimation biases from serological survey data. AG# = survey defined age-grouping scheme # (16 total AG assessed), ReproAG# = survey defined age-grouping scheme among reproductive ages only # (5 total ReproAG assessed).

<table>
<thead>
<tr>
<th>Step #</th>
<th>Type of Biases Examined</th>
<th>Description of Bias Analysis</th>
<th>Analytic Methods</th>
<th>Age Grouping (AG) Schemes</th>
<th>Sample Sizes per age group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step #1</td>
<td>Analytic Method</td>
<td>Compared the true parameter values to estimated parameters per analytic method based on perfect seroprevalence data at every age in years. The results are referred to as ‘best analytic estimates’</td>
<td>Fractional Polynomials (FP), Penalized Regression Splines (PRS), Piecewise Constant (PC)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Step #2</td>
<td>Analytic Method</td>
<td>Compared the true parameter values and best analytic estimates to estimated parameters per analytic method based on perfect seroprevalence data of each defined age-group for all Age Grouping (AG) schemes</td>
<td>Fractional Polynomials (FP), Penalized Regression Splines (PRS), Piecewise Constant (PC)</td>
<td>AG1-AG16, ReproAG1-ReproAG5</td>
<td>NA</td>
</tr>
<tr>
<td>Step #3</td>
<td>Analytic Method Bias</td>
<td>Compared the true parameter values and best analytic estimates to estimated parameters per analytic method based on simulated seroprevalence data by population sampling per Age Grouping (AG) scheme, and per sample size</td>
<td>Fractional Polynomials (FP), Penalized Regression Splines (PRS), Piecewise Constant (PC)</td>
<td>AG13-AG16</td>
<td>1000, 250, 60</td>
</tr>
</tbody>
</table>

3.3.1 Simulation of Three Endemic Rubella Populations

The first step was to simulate realistic populations with endemic rubella (i.e., no vaccination). The epidemiology of rubella is broadly known. Estimates of $R_0$ for rubella based on empirical data yield a median value of 5.2 (90% CI 4.0-6.7) ([Lessler and Metcalf, 2013](#), although estimates range from 3 in Europe to 12 in Ethiopia ([Anderson and May, 1991](#), [Cutts et al., 2000](#), [Edmunds et al., 2000](#)). Generally, the average age of infection for rubella ranges between 7 and 9 years old, and force of infection peaks between ages 5 and 10 ([Anderson and May, 1991](#), [Grenfell and...](#))
We simulated typical rubella transmission dynamics in endemic settings based on these known parameters using an established deterministic age-structured mathematical model, introduced in Metcalf et al. (2012b) and Metcalf et al. (2012a). The key feature of the model was a matrix that at every discrete time-step defines transitions from four possible epidemiological stages (e.g., maternally immune, susceptible, infected, and recovered) and age combinations to every other possible epidemiological stage and age combination. The time-step was set to two weeks roughly based on the generation time of rubella, thus there are 24 discrete opportunities to move in and out of epidemiological stages per year.

We simulated rubella dynamics for three populations based on different values of $R_0$ (3, 5, and 7). We assumed the same population age structure, age-specific fertility rates, and birth rate for each population based on the region of rural Bihar in India (Office of Registrar General & Census Commissioner, 2012; Office of the Registrar General & Census Commissioner, 2011). We chose a Bihar-like population because, like the majority of countries that have not yet introduced the rubella vaccine, Bihar has a growing population. The value of the assumed $R_0$ combined with the population size $N$ sets the intensity of transmission with $R_0 = \bar{\beta} N$, where $\bar{\beta}$ is average transmission. The model assumed seasonal forcing, where the intensity of transmission varied over the year (affecting all age-groups the same), by setting $\beta_t = \bar{\beta}(1 - \alpha \cos(2\pi t/24))$ for each time step $t$; $\alpha$ was set to 0.35 which reflects the factor by which the peak and lowest value of $R_0$ differs from the average. We also assumed a pattern of contact over age proportional to what was measured in Great Britain via the POLYMOD diary study (Mossong et al., 2008). Population survival rates per age class were estimated by matching the region’s life expectancy of 65.6 to a United Nations model life table for South Asia (Office of the Registrar General & Census Commissioner, 2012b; UN Population Division, 2010).

We simulated each population for 150 years to allow both infectious disease and population equilibrium behavior to be reached. We then rescaled the age transmission rates matrices, i.e., the who acquires infection from whom (WAIFW), such that they reflected the assumed $R_0$. 

(Anderson, 1985).
We then simulated the populations for another 90 years until epidemiological stability was once again established. Epidemiological and demographic stability means that the age structure of the population is stable over time, rubella outbreaks are consistent in size and occur cyclically based on the assumed cosine function representing seasonality, and the mean ASFOI across one full cyclical phase is stable (John, 1990a). Whitaker and Farrington (2004) analyzed rubella dynamics in the U.K. and found that cyclical outbreaks of rubella made little difference to the ASFOI averaged over the length of the phase. As a result, seasonality can generally be ignored, and we can assume a time invariant ASFOI. To facilitate ease of interpretation of results and identification of drivers of key biases, we assume a stable population with a constant ASFOI. We review the consequences of a time invariant ASFOI on parameter estimation in the discussion section. Figure 3.1 displays the stable age structure for all three populations with a stable annual growth rate of 1.72%. The associated epidemiological equilibriums based on the three different assumed values of $R_0$ (3, 5, and 7) are shown in Figure 3.2. Seroprevalence steeply rose from age 1 years old to approximately 15 years old, and then gradually climbed to 100% after 15 years old. The average age of infection ranged from 5 to 10 years old (Figure 3.2). As a result of the assumed pattern of contact over age (Mossong et al., 2008), the three different ASFOI curves peaked at approximately age 10 years old and again at approximately age 37 years old due to high assumed contact rates between children and parents (Figure 3.2).

Each simulation provided the number of individuals at every age and epidemiological state. Individuals in the maternally immune, infected, or recovered epidemiological stages were classified as seropositive, and individuals in the susceptible stage were classified as seronegative. Based on each survey design designation of age-groups, discussed below, we estimated the proportion seropositive by age-group either by assuming perfect information of seroprevalence, or by simulating serosurveys by sampling a finite number of individuals from the population. We then used these age seroprevalence data to estimate key epidemiological parameters. The age-structured model is presented in detail in Metcalf et al. (2012b) and Metcalf et al. (2012a), and are therefore not further discussed here.
Figure 3.1: Population age structure for male and female (M/F) populations alike. All three explored populations had the same demographic equilibrium with a growth rate of 1.72%.

Figure 3.2: Population age seroprevalence and age-specific force of infection. Each explored population had a different epidemiological equilibrium based on the assumed value of $R_0$. a) True proportion seropositive by age for three populations with different assumed $R_0$ values of 3, 5, and 7. The grey lines are the average ages of infection that correspond with the line style of each population. b) True force of infection by age for the three populations with different assumed values of $R_0$. 
3.3.2 Serological Survey Design

Age-related changes in the force of infection are well established and generally attributed to changes in the pattern of contact over age as captured in our simulations (Grenfell and Anderson, 1985; Mossong et al., 2008; Anderson and May, 1983, 1985). As a result, the chosen age-groups of a survey design will affect estimates of the force of infection, \( R_0 \), and the CRS incidence rate. We explored 21 different survey defined age-grouping (AG) schemes. Literature on past rubella serological surveys, reviewed in depth by Goodson et al. (2011), Thompson and Odahowski (2016), and Vynnycky et al. (2016), reveals that age-groups of rubella serological survey designs have little consistency, but two general trends emerge: many rubella serological surveys only sample from women of reproductive ages, and five year age-groups are favored. Therefore, we assessed five AG schemes that focused solely on reproductive ages (ReproAG); and 16 AG schemes across all ages generally based around increments of five year age-groups. Survey defined age-group bias was systematically examined in four stages based on the shape of the ASFOI curve for ease of presenting and interpreting results: groupings between ages 1 and 10, groupings between ages 10 and 20, groupings between ages 20 and 50, and reproductive age-groupings between 15 and 50.

After assessing biases from survey defined AG schemes based on perfect seroprevalence data by age-group, we chose four final AG schemes to simulate serological surveys, selected to reflect the fewest number of age-groups that resulted in the smallest bias.

A review of the literature cited by Goodson et al. (2011), Thompson and Odahowski (2016) and Vynnycky et al. (2016), also showed that there is little consistency in sample sizes for age-groups in past rubella serological surveys. The number of samples to be drawn for each age-group in an age-stratified serological survey must be sufficiently large to detect differences in the proportion seropositive between age-groups. This number is primarily determined by the true proportion seropositive for each age-group (Daniel and Cross, 2013). In realistic public health settings, we do not know the details of age-immunity profiles prior to conducting the serological sample, and are consequently unable to designate specific sample sizes to each age-group. The same number of samples is therefore sampled for each designated age-group. The mean sample size across the
full range of proportion seropositive (0-1), given 95% precision and an \( \alpha \) of 0.05 is approximately 1000 samples. Assuming 90% precision the mean sample size is approximately 250. Assuming 80% precision the mean sample size is approximately 60. \cite{Daniel2013, Lwanga1991}. We ran simulations for three sample sizes per age-group (1000, 250, and 60).

### 3.3.3 Simulated Serological Surveys

We simulated 100 age-stratified serological surveys for each population and survey design to allow for variation in binomial sampling; for a total of 3,600 serological surveys (3 populations \( \times 4 \) age-grouping schemes \( \times 3 \) sample sizes \( \times 100 \) serosurveys). To reflect the sampling process of a serological survey, we created a vector equivalent to the length of the total population at the time of the serosurvey for each age-group. The vector identified each individual by their serostatus (1=seropositive, 0=seronegative). We then drew a random sample of individuals from this vector for each age group without replacement until reaching the sample size per age group set by the serological sampling scheme for each survey design. For simplicity, we assumed perfect sensitivity and specificity of a rubella IgG antibody diagnostic test in our simulated survey. For both our full range of age-groups where we assumed perfect information (i.e., no sampling process), and for each simulated serological survey (i.e., population sampling as described above), we estimated epidemiological variables of interest: seroprevalence by age, the ASFOI, the \( R_0 \) of rubella, and the incidence rate of CRS.

### 3.3.4 Estimation of Age Specific Force of Infection - Three Analytic Approaches

The Age-Specific Force of Infection (ASFOI) is the age-specific rate at which susceptible individuals acquire infection. We used three different analytic approaches to extract rubella age-specific force of infection from age seroprevalence data: 1) parametric fractional polynomials, 2) semiparametric
penalized regression splines, and 3) two-age-group piecewise constants. We used both parametric and semiparametric approaches to demonstrate the range of analytic procedures, and we additionally chose WHO’s analytic method of choice (the two-age-group piecewise constant approach) to evaluate current estimates of CRS incidence rates across the world. See Hens et al. (2010) for a succinct discussion of additional seroprevalence data analytic methods not selected for this analysis.

In principle for a fully immunizing infection such as rubella, serological data of proportion seropositive in each age-group corresponds directly to the cumulative proportion of infected individuals of each age-group \( a \), assuming that the sample collected represents a typical birth cohort, ignoring vaccination, and assuming 100% sensitivity in serological assays. The methods discussed in detail below were built from a rich foundation of work by Griffiths (1974), Grenfell and Anderson (1985), and specifically Muench (1959), who first proposed to model seroprevalence by age (\( \pi(a) \)), as

\[
\pi(a) = k(1 - \exp(-\lambda a)),
\]

where (1 – \( k \)) is the proportion of the population that will remain susceptible, and \( \lambda \) is the age constant force of infection. In addition to original works, we relied upon the Hens et al. (2012) textbook and online supplemental material for descriptions of the chosen analytic methods.

Fractional polynomial and penalized regression spline methods derive the ASFOI (\( \lambda(a) \)) from the seroprevalence by age (\( \pi(a) \)). The probability of seroprevalence by age (\( \pi(a) \)) was modeled as

\[
g(P(Y = 1|a)) = g(\pi(a)) = \eta(a),
\]

where \( \eta(a) \) was the linear predictor, and \( g \) is the link function. We explored two types of link functions for binary seroprevalence data, the logit link function and the complementary log-log link function. \( \eta \) took the form of a fractional polynomial function or spline function below.
Fractional polynomials, as introduced by Royston and Altman (1994), based on the work on higher-order polynomials by Grenfell and Anderson (1985), allow for a large range of relationships between age and the force of infection from constant to flexible curve shapes, while maintaining the attractive features of parametric models. A fractional polynomial of degree $m$ for the linear predictor is defined as

$$\eta_m(a, \beta, p_1, p_2 \ldots p_m) = \sum_{i=0}^{m} \beta_i H_i(a), \quad (3.3)$$

where $m$ is an integer, $p_1 \leq p_2 \leq \ldots \leq p_m$ is a sequence of powers, and $H_i(a)$ is a transformation given by

$$H_i(a) = \begin{cases} 
  a^{p_i} & \text{if } p_i \neq p_{i-1}, \\
  H_{i-1}(a) \times \log(a) & \text{if } p_i = p_{i-1},
\end{cases} \quad (3.4)$$

with $p_0 = 0$, $H_0 = 1$, and $a^{p_i} = \log(a)$ if $p_i = 0$. In our final analysis, we used a logit link function, such that the proportion seropositive ($\pi(a)$) is defined as

$$\pi(a) = \frac{e^{\eta(a)}}{1 + e^{\eta(a)}}. \quad (3.5)$$

The results were robust to the use of the complementary log-log link function as well.

Our analysis explored only first ($m = 1$) and second degree ($m = 2$) polynomials. The search for the ‘best’ fractional polynomial model was a two-step process. First, we fit first and second degree polynomial models by estimating values of the power sequence $(p_1, p_m)$ between [-2, 3] that resulted in the lowest deviance as suggested by Hens et al. (2012). We used a step size of 0.01 to explore power values between [-2, 3] for the first degree polynomial models, and a step size of 0.1 to explore power values between [-2, 3] for the second degree polynomial models, as suggested by Hens et al. (2012). The baseline deviance comes from the null model of $\eta_1(a, \beta, 1)$. The power values were constrained to comply with a monotonically increasing seroprevalence model assumption, such that $\eta'_m(a, \hat{\beta}, p) \geq 0$. The second step was to chose between the first and second degree polynomial models using a chi-square test. The second degree polynomial model was chosen if $D(1, \tilde{p}) - D(2, \tilde{p}) > \chi^2_{2, 0.9}$, where $\tilde{p}$ is the power sequence for the model that has the best goodness
of fit among models of the same degree; otherwise, the first degree polynomial model was chosen (Royston and Altman, 1994). To prevent extreme estimates at the oldest and youngest ages, we assumed that 99% of the population at age 6 months old was seronegative (i.e., susceptible to rubella infection), as has been shown empirically in endemic rubella settings (Waaijenborg et al., 2013), and we assumed that 99% of the population at age 90 years old was seropositive.

We additionally estimated ASFOI from seroprevalence data using a highly flexible semiparametric model with penalized regression splines. Penalized regression splines smooth over a scatterplot based on a defined spline basis and penalty that controls ‘wiggliness’ by measuring tradeoffs between smoothness and fit to the data. The linear predictor for a penalized regression spline was defined as

$$\eta(a) = f(a), \quad (3.6)$$

where $$f(a)$$ is a smooth function of the age covariate, $$a$$. Penalized regression splines are estimated by first choosing a spline basis (and associated penalty) and basis dimension, and then using cross validation to estimate the smoothing parameter, and penalized least squares to estimate the regression splines (Wood, 2006). We selected the ‘best’ penalized regression spline between two spline basis functions (cubic regression spline, and thin plate regression spline) and two link functions (logit and complementary log-log), by minimizing Bayesian Information Criteria (BIC). If the link function was logit, the proportion seropositive ($$\pi(a)$$) is defined in equation 3.5. If the link function was complementary log-log, the proportion seropositive ($$\pi(a)$$) was defined as

$$\pi(a) = 1 - e^{-e^{\eta(a)}}. \quad (3.7)$$

The basis dimension simply designates an upper bound on the flexibility of the fit. In contrast, the smoothing parameter, which is estimated by minimizing the generalized cross validation score, controls the actual degrees of freedom and defines the ‘wiggliness’ penalty in the least squares fitting (Wood, 2006). The choice of the basis dimension was large enough to have enough degrees of freedom to represent the underlying data (Wood, 2006). The ‘mgcv’ library in R (Wood, 2006; R Core Team, 2016) was chosen to model the penalized regression splines for its computational
efficiency, automated selection of the smoothness parameter, and goodness of fit to the data.

For both analytic models, fractional polynomials and penalized regression splines, the force of infection by age \( \hat{\lambda}(a) \) can be obtained from the estimated proportion seropositive by age \( \hat{\pi}(a) \) by

\[
\hat{\lambda}(a) = \frac{\hat{\pi}'(a)}{(1 - \hat{\pi}(a))}.
\]  \hspace{1cm} (3.8)

The last analytic approach we explored to estimate the age-specific force of infection was a two-age-group piecewise constant model (i.e., separate estimates for each age-group with no explicit parametric form). Fitting piecewise constant force of infections to seroprevalence data is a popular approach \cite{Hens2010}, and is the method used to estimate the global CRS burden estimates for the WHO \cite{Vynnycky2016}. As such, we mimicked the method used by Vynnycky et al. \cite{Vynnycky2016} (model B), which assumed two-age-groups and 100% sensitivity of diagnostic assays (i.e., no false negatives). The piecewise constant hazard of infection for ‘younger’ individuals (ages 0.5 to 13 years old) was represented by \( \bar{\lambda}_y \), and the piecewise constant hazard of infection for ‘older’ individuals (ages 13 to 50 years old) was represented by \( \bar{\lambda}_o \). The relationship between the proportion seronegative by age \( a \) \((1 - \pi(a))\), and the older and younger age forces of infection (\( \bar{\lambda}_o \) and \( \bar{\lambda}_y \)) per Muench \cite{Muench1959} was

\[
1 - \pi(a) = \begin{cases} 
\ e^{\bar{\lambda}_y(a-0.5)} & a < 13\text{years} \\
\ e^{-12.5\bar{\lambda}_y a - 13} & a \geq 13\text{years}. 
\end{cases}
\]  \hspace{1cm} (3.9)
Integrating this expression over the age-group of interest ($a_j$ to $a_k$) results in the proportion seronegative for each age-group ($1 - \pi(A_{j,k})$) defined below

$$1 - \pi(A_{j,k}) = \begin{cases} 
  \frac{e^{-\bar{\lambda}_y(a_j-0.5)} - e^{-\bar{\lambda}_y(a_k-0.5)}}{\bar{\lambda}_y(a_k - a_j)} & a_j, a_k < 13\text{years} \\
  \frac{(13 - a_j)s(A_{j,13}) + (a_k - 13)s(A_{13,k})}{a_k - a_j} & a_j < 13\text{years, } a_k \geq 13\text{years} \\
  \frac{e^{-12\bar{\lambda}_y(e^{-\bar{\lambda}_y(a_j-13)} - e^{-\bar{\lambda}_o(a_k-13)})}}{\bar{\lambda}_o(a_k - a_j)} & a_j, a_k \geq 13\text{years}.
\end{cases} \quad (3.10)$$

The ASFOI was then be estimated by identifying the value of $\bar{\lambda}_y$ and $\bar{\lambda}_o$ that maximizes the binomial likelihood of seronegative individuals by age-group $A_{j,k}$. Optimization of the likelihood was conducted using the Nelder-Mead algorithm in the optim function the R base library [R Core Team, 2016]. The estimated proportion seropositive ($\hat{\pi}(a)$) was then back estimated for all ages based on $\bar{\lambda}_y$ and $\bar{\lambda}_o$ defined as

$$\hat{\pi}(a) = \begin{cases} 
  1 - (e^{-\bar{\lambda}_y(a-0.5)}) & a < 13\text{years} \\
  1 - (e^{-12.5\bar{\lambda}_y}e^{-\bar{\lambda}_o(a-13)}) & a \geq 13\text{years}.
\end{cases} \quad (3.11)$$

To avoid biases resulting from maternal immunity, we excluded data of individuals aged less than one year in all analytic methods. This deletion encompassed the range of individuals protected by maternal immunity for rubella [Anderson and May 1991, Grenfell and Anderson 1985, Anderson and May 1983]. Two known biases arise from the interpretation of serological data when estimating forces of infection: seropositivity may decrease with age due to waining antibody levels, and some seronegative individuals may be immune [Farrington 1990]. In this analysis, we ignored these biases by assuming that antibody tests have 100% sensitivity and specificity.
3.3.5 Estimation of CRS Incidence Rate and $R_0$

Estimation of the CRS incidence rate ($CRSIR$) was limited to inference based on the estimated proportion seropositive by age and the ASFOI. The estimated CRS incidence per 100,000 live births for each reproductive age in years ($a$) was defined as

$$\hat{CRSIR}_a = (1 - \hat{\pi}(a)) \times (1 - e^{-16\hat{\lambda}(a)/52}) \times 0.65 \times 100,000,$$

where $(1 - \hat{\pi}(a))$ is the estimated probability of being seronegative at age $a$, and $\hat{\lambda}_a$ is the estimated force of infection at age $a$. We also assumed that 65% of babies born to mothers infected during the first 16 weeks of pregnancy were born with CRS ([Vynnycky et al., 2003]). The population CRS incidence per 100,000 live births was calculated as the weighted mean of CRS incidence per 100,000 live births in each reproductive age in years ($a$), defined as

$$\hat{CRSIR} = \sum_{a=15}^{45} \hat{CRSIR}_a \times \left( \frac{b_a}{\sum_{a=15}^{45} b_a} \right),$$

where $b_a$ is the number of births to women of age $a$.

The basic reproduction number ($R_0$) is defined as the expected number of secondary infections after the introduction of a typical infectious individual in a completely susceptible population. Estimating $R_0$ is complex in the case of an age-varying force of infection. Next generation methods indicate that $R_0$ is defined as the dominant eigenvalue of the next generation matrix ($M$) ([Diekmann et al., 1990]), i.e.,

$$M_{a,j} = W_{a,j}S_a.$$

Here, the matrix $W_{a,j}$ is also known as the Who-Acquires-Infection-From-Whom (WAIFW) matrix, which contains the values of the transmission rates between ages. The element in the $a^{th}$ row and $j^{th}$ column denotes the probability that an infected individual age $j$ will infect a susceptible individual age $a$ per one generation of the infection. The vector $S_a$ is the number of expected susceptible individuals of age $a$ in a completely susceptible population ([Diekmann et al., 1990]). To link this
back to the ASFOI ($\hat{\lambda}_a$), we assumed a known matrix defining the pattern of contact between ages $a$ and $j$ ($\beta_{a,j}$); here we used the POLYMOD diary study contact matrix based on Great Britain contact data, as was used for the simulation [Mossong et al., 2008]. The WAIFW ($W_{a,j}$) is the product of the contact matrix ($\beta_{a,j}$) and some constant $c$, which scales $\beta_{a,j}$ to reflect transmission given that contact has occurred. We know then that the FOI at every age $a$ should be defined by

$$\bar{\lambda}_a = \sum_{j=1}^{n} \hat{c} \beta_{a,j} \frac{I_j}{N},$$

where $\beta_{a,j}$ is the vector of contacts incurred by individuals in age $a$ with individuals in age $j$, $I_j$ is the number of infected individuals age $j$ and $N$ is the total population size (because we assumed frequency dependent transmission). With knowledge of $\beta$, $I_j$, and $N$ we can therefore identify $\hat{c}$ such that the distance between $\hat{\lambda}_a$ and $\bar{\lambda}_a$ is minimized. The estimate of $R_0$ was then calculated as described above. It is reasonable to assume that the total population size is known, and sensitivity to patterns of contact over age can be assessed; but age incidence data is rarely available. For the sake of this analysis, we assumed complete knowledge of $\vec{S}_a$, $\beta_{a,j}$, and $\frac{I_j}{N}$. We were specifically interested in exploring $R_0$ estimation bias from seroprevalence data and inferred ASFOI data, and error in these parameters would confound our understanding of the effect. All statistical analyses were conducted using R 3.2.2 [R Core Team, 2016].

3.4 Results

We systematically explored biases in parameter estimation by using a stepwise approach described above and in Table 3.2. The following sections report findings from each bias analysis: i) Analytic Method Bias, ii) Survey Defined Age Group Bias, and iii) Sampling Bias.
3.4.1 Analytic Method Bias

We first looked at analytic method bias by providing each analytic method with perfect seroprevalence data at every age in years to estimate seroprevalence by age, the ASFOI, CRS incidence per 100,000 live births, and the \( R_0 \) of rubella. We refer to estimates from this finely resolved age data as ‘best analytic estimates.’ Figure 3.3 shows the results for the population with an assumed \( R_0 \) of 5. As seen by comparing Figures 3.3a-c, Figure 3.3b (the penalized regression splines analytic method) best captured the proportion seropositive and subsequent ASFOI, as the estimated blue line falls on top of the true red line at almost every age.

The fractional polynomial parametric model (Figure 3.3a) followed general trends in the data, i.e., a steep increase in the proportion seropositive until approximately 10 years old, followed by a slow increase up to 50 years old. However, the fractional polynomial model was unable to capture small changes in the proportion seropositive curve due to the constraints of the model form. As a result, estimates undulated between over and underestimating the proportion seropositive by age, and the ASFOI. Small differences in the estimated proportion seropositive using the fractional polynomial method resulted in large differences in the estimated ASFOI.

Figure 3.3c shows the estimated proportion seropositive and ASFOI using the piecewise constant approach for two age-groups (13,50]. The two-age-group force of infection lines oversimplified the true age dependent force of infection, which resulted in a concave shape between ages 0 and 20 years old, and a convex function between ages 10 and 38 years old. Jensen’s inequality theorem proves that non-linear averaging can result in underestimates or overestimates depending on the shape of the function (Jensen, 1906). Therefore, if \( X \) is a random variable and \( \varphi \) is a concave function, then:

\[
\varphi(\mathbb{E}(X)) \geq \mathbb{E}(\varphi(X)) \quad (3.16)
\]

If \( \varphi \) is a convex function, then:

\[
\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X)) \quad (3.17)
\]
The average of the concave function between ages 1 and 15 years in the ASFOI of rubella underestimated the function of the average; and the average of the convex function between ages 13 and 38 years in the ASFOI of rubella overestimated the function of the average (Figure 3.3c). Additional two-age-group force of infection estimates of (10,50] and (15,50] are shown in dark and light blue, respectively, in Figure 3.6. As we increased the size of the first age-group (qualitatively similar to shifting the ASFOI curve to the left), the less the average of the older age FOI, $\bar{\lambda}_o$, function differed from the function of the average, and the more the estimates moved closer to the true value of the CRS incidence rate.

Given the accuracy of the ASFOI estimates using penalized regression splines, CRS incidence rate by age (Figure 3.3d), population CRS incidence rate (Figure 3.3e), and $R_0$ (Figure 3.3f) were also best estimated with this analytic method. The fractional polynomial and piecewise constant methods overestimated population CRS incidence rate and underestimated $R_0$ (Figure 3.3e & f). The effect of ASFOI bias on $R_0$ was fairly straightforward in our analysis. We assumed perfect information of $\tilde{S}_a$, $\beta_{a,j}$, and $I_t/N$ (see equations 3.14 and 3.15); an underestimate of $\hat{\lambda}_a$ means an underestimate of the constant $\hat{c}$, and therefore an underestimate in $R_0$. This means the ASFOI and $R_0$ are positively correlated, and for the two-age-group piecewise constant method, the younger age FOI ($\hat{\lambda}_y$) appears to play the prominent role in determining $R_0$. The fractional polynomials method underestimates $R_0$ by 0.33 (i.e., 7%), and the piecewise constant method underestimates $R_0$ by 0.24 (i.e., 5%) (Figure 3.3f).

The effect of the ASFOI estimation bias on the population CRS incidence rate is more complicated because it was estimated as an average of the CRS incidence rate over reproductive ages, weighted by births per age. The fractional polynomial method overestimated the CRS incidence rate by 12 cases per 100,000 births, (i.e., 23%), and the piecewise constant method overestimated the CRS incidence rate by 15 cases per 100,000 births (i.e., 27%) (Figure 3.3d-e). The resulting bias was dependent on the age the CRS incidence rate shifted from overestimating to underestimating CRS incidence rates across reproductive ages, and on the age-specific fertility rates. For the
majority of countries where rubella containing vaccine has not been introduced (World Health Organization, 2016c), most births occur between ages 20 and 30 years old (United Nations, 2015), i.e., the ages where both fractional polynomials and piecewise constant approaches were generally overestimating the force of infection.

Our general finding that the penalized regression splines method correctly estimated parameters, and that the fractional polynomials and two-age-group piecewise constant methods overestimated the CRS incidence rate and underestimated $R_0$, additionally held true for other populations with assumed $R_0$ values of 3 and 7 (Figures 3.4 and 3.5). However, the magnitudes of biases for the fractional polynomials and piecewise constant methods was different across populations. For the population with an assumed $R_0$ of 3, fractional polynomials underestimated $R_0$ by 8% and overestimated the CRS incidence rate by 34%, and two-age-group piecewise constants underestimated $R_0$ by 9% and overestimated the CRS incidence rate by 38%. For the population with an assumed $R_0$ of 7, fractional polynomials underestimated $R_0$ by 6% and overestimated the CRS incidence rate by 14%, and two-age-group piecewise constants underestimated $R_0$ by 2% and overestimated the CRS incidence rate by 10%.

We additionally explored analytic method bias if only reproductive ages (15 to 50 years old) were sampled, a common survey design for rubella serological surveys (Vynnycky et al., 2016). We supplied each analytic method with perfect seroprevalence data for reproductive ages in years (15 to 50 years old) in order to estimate seroprevalence by age, the ASFOI, and CRS incidence per 100,000 live births. Figure 3.7 shows the results for the population with assumed $R_0$ value of 5. In the fractional polynomials analysis, we assumed that 99% of the population was susceptible at age 6 months, therefore the estimated proportion seropositive curve starts at 0.01 at age 6 months. Using the fractional polynomial method, we found that the the CRS incidence rate estimation was more accurate when only using reproductive-age serological data as compared to using all-age serological data, and this result held for the other populations as well (Figures 3.8 and 3.9). Specifically, the CRS incidence rate was overestimated between 5 to 7%, across
Figure 3.3: Epidemiological parameter estimates from all-age seroprevalence data on the population with an assumed \( R_0 \) of 5: Estimation of seroprevalence by age and the ASFOI based on perfect seroprevalence data at every age in year for the a) fractional polynomials (FP), b) penalized regression splines (PRS) and c) piecewise constant (PC) analytic methods. In figures a-c, the red solid lines represent the true proportion seropositive and the ASFOI, while the non-solid lines represent the estimated parameters. d) The estimated CRS incidence rates by age, the solid red line is the truth, and the non-solid lines are estimates per analytic method: Fractional Polynomials (FP), Penalized Regression Splines (PRS), and Piecewise Constants (PC). e) Estimated average population CRS incidence per 100,000 live births (weighted by births per age) per analytic technique. The true CRS incidence rate is represented as the vertical red line at 52.4 CRS cases per 100,000 live births. f) Estimated \( R_0 \) for each analytic method, the true \( R_0 \) is displayed as the vertical red line at 5.
Figure 3.4: Epidemiological parameter estimates from all-age seroprevalence data on the population with an assumed $R_0$ of 3: Estimation of seroprevalence by age and the ASFOI based on perfect seroprevalence data at every age in year for the a) fractional polynomials (FP), b) penalized regression splines (PRS) and c) piecewise constant (PC) analytic methods. In figures a-c, the red solid lines represent the true proportion seropositive and the ASFOI, while the non-solid lines represent the estimated parameters. d) The estimated CRS incidence rates by age, the solid red line is the truth, and the non-solid lines are estimates per analytic method: Fractional Polynomials (FP), Penalized Regression Splines (PRS), and Piecewise Constants (PC). e) Estimated average population CRS incidence per 100,000 live births (weighted by births per age) per analytic technique. The true CRS incidence rate is represented as the vertical red line at 103.5 CRS cases per 100,000 live births. f) Estimated $R_0$ for each analytic method, the true $R_0$ is displayed as the vertical red line at 3.
Figure 3.5: Epidemiological parameter estimates from all-age seroprevalence data on the population with an assumed $R_0$ of 7. Estimation of seroprevalence by age and the ASFOI based on perfect seroprevalence data at every age in year for the a) fractional polynomials (FP), b) penalized regression splines (PRS) and c) piecewise constant (PC) analytic methods. In figures a-c, the red solid lines represent the true proportion seropositive and the ASFOI, while the non-solid lines represent the estimated parameters. d) The estimated CRS incidence rates by age, the solid red line is the truth, and the non-solid lines are estimates per analytic method: Fractional Polynomials (FP), Penalized Regression Splines (PRS), and Piecewise Constants (PC). e) Estimated average population CRS incidence per 100,000 live births (weighted by births per age) per analytic technique. The true CRS incidence rate is represented as the vertical red line at 29.2 CRS cases per 100,000 live births. f) Estimated $R_0$ for each analytic method, the true $R_0$ is displayed as the vertical red line at 7.
Figure 3.6: Two-age-group piecewise constant estimation of the ASFOI across three, two-age-groups.

populations when using only reproductive-age serological data, as compared to a 14 to 24% overestimation when using all-ages serological data. Using a penalized regression splines analysis, we found that CRS incidence rate estimates were minimally biased away from the true parameter value using only reproductive-age data. For populations with an assumed $R_0$ values of 3, 5, and 7, the CRS incidence rates were overestimated by 1, 3, and 5% respectively (Figure 3.7, 3.8, and 3.9).

Interestingly, we found that when using the two-age-group piecewise constant method, the younger age FOI ($\hat{\lambda}_y$) was overestimated in the absence of younger age seroprevalence data and the older age FOI ($\hat{\lambda}_o$), was underestimated, which is in direct contrast to the direction of bias using all-age data (Figure 3.7). Excluding ages 13 and 14 years old from the data was similar to shifting the ASFOI curve to the left, i.e., shifting the lower age bound up as seen in Figure 3.6. The force of infection between ages 15 and 50 looked more like a cubic function, and resulted in
Figure 3.7: Epidemiological parameter estimates from reproductive-age seroprevalence data on the population with an assumed \( R_0 \) of 5: Estimation of seroprevalence by age and the ASFOI based on perfect data at each reproductive age in years 15 to 50 for the a) fractional polynomials (FP), b) penalized regression splines (PRS) and c) piecewise constant (PC) analytic methods. In figures a-c, the red solid lines represent the true proportion seropositive and the ASFOI, while the non-solid lines represent the estimated parameters. d) The estimated CRS incidence rates by age, the solid red line is the truth, and the non-solid lines are estimates per analytic method: Fractional Polynomials (FP), Penalized Regression Splines (PRS), and Piecewise Constants (PC). e) Estimated average population CRS incidence per 100,000 live births (weighted by births per age) per analytic technique. The true CRS incidence rate is represented as the vertical red line at 52.4 CRS cases per 100,000 live births. f) Estimated \( R_0 \) for each analytic method, the true \( R_0 \) is displayed as the vertical red line at 5.

an underestimate in the older age FOI (\( \hat{\lambda}_o \)), and an underestimate of the CRS incidence rate below the true parameter value. Different populations with different true CRS incidence rates determined the magnitude of underestimation: for the population with an assumed \( R_0 \) of 3, the CRS incidence rate was underestimated by up to 5%; for the population with an assumed \( R_0 \) of 5, the CRS incidence rate was underestimated by up to 10%; and for the population with an assumed \( R_0 \) of 7, the CRS incidence rate was underestimated by up to 14% (Figure 3.7, 3.8 and 3.9).
Figure 3.8: Epidemiological parameter estimates from reproductive-age seroprevalence data on the population with an assumed $R_0$ of 3: Estimation of seroprevalence by age and the ASFOI based on perfect data at each reproductive age in years 15 to 50 for the a) fractional polynomials (FP), b) penalized regression splines (PRS) and c) piecewise constant (PC) analytic methods. In figures a-c, the red solid lines represent the true proportion seropositive and the ASFOI, while the non-solid lines represent the estimated parameters. d) The estimated CRS incidence rates by age, the solid red line is the truth, and the non-solid lines are estimates per analytic method: Fractional Polynomials (FP), Penalized Regression Splines (PRS), and Piecewise Constants (PC). e) Estimated average population CRS incidence per 100,000 live births (weighted by births per age) per analytic technique. The true CRS incidence rate is represented as the vertical red line at 52.4 CRS cases per 100,000 live births. f) Estimated $R_0$ for each analytic method, the true $R_0$ is displayed as the vertical red line at 3.
Figure 3.9: Epidemiological parameter estimates from reproductive-age seroprevalence data on the population with an assumed $R_0$ of 7. Estimation of seroprevalence by age and the ASFOI based on perfect data at each reproductive age in years 15 to 50 for the a) fractional polynomials (FP), b) penalized regression splines (PRS) and c) piecewise constant (PC) analytic methods. In figures a-c, the red solid lines represent the true proportion seropositive and the ASFOI, while the non-solid lines represent the estimated parameters. d) The estimated CRS incidence rates by age, the solid red line is the truth, and the non-solid lines are estimates per analytic method: Fractional Polynomials (FP), Penalized Regression Splines (PRS), and Piecewise Constants (PC). e) Estimated average population CRS incidence per 100,000 live births (weighted by births per age) per analytic technique. The true CRS incidence rate is represented as the vertical red line at 52.4 CRS cases per 100,000 live births. f) Estimated $R_0$ for each analytic method, the true $R_0$ is displayed as the vertical red line at 7.
3.4.2 Survey Defined Age Group Bias

We systematically explored analytic methods and survey design defined age-group bias by providing each analytic model with perfect seroprevalence data by age-group for 21 Age-Grouping (AG) schemes, and then estimated seroprevalence by age, the ASFOI, CRS incidence per 100,000 live births, and the $R_0$ of rubella. Perfect seroprevalence data by age-group consisted of the average proportion seropositive per defined age-group assigned to the mid-age of the group. We compared epidemiological parameter estimates to the true parameter values, and also compared estimates to the best analytic estimates using perfect all-age data as assessed in the previous section. Given the accuracy of the best analytic estimates using the penalized regression splines, biases from grouping defined ages will always move away from both the true parameter value and the best analytic estimate. In contrast, best analytic parameter estimates using fractional polynomials and piecewise constant methods were biased even with finely resolved age seroprevalence data. Therefore, the direction of bias from grouping defined ages can either move away from both the true parameter and the best analytic estimates, or away from the best analytic estimate and towards the true parameter values. We discuss examples of both these directions of bias using fractional polynomials and piecewise constant analytic approaches below.

We examined 21 age-grouping schemes in four stages: grouping between ages 1 and 10, groupings between ages 10 and 20, groupings between ages 20 and 50, and groupings among reproductive ages 15 and 50.

The effect of age-grouping on the estimation of epidemiological parameters depends foremost on the age structure of the population, i.e., the growth rate in a stable population. Consider a population with a rectangle population structure where there is a constant birth rate and no one dies until some age, when everyone dies. In this population, age-based sampling will draw equally from every age in a designated age-group, and therefore the mid-age of any age-group is equal to the average age of individuals from a perfectly random sampled age-group. In contrast, in a
population with a positive growth rate, such as the populations we have simulated (Figure 3.1), age-based samples will be skewed to the younger ages in a designated age-group. The mid-age of the age-group will be larger than the average age of individuals from the age-group sampled, and this discrepancy widens as the population growth rate increases. This is a basic demographic concept that can be resolved by age weighting samples to control for population structure, or by estimating the mid-age as the average sampled age of the age-group if the raw data is available. However, this is often not done, and we will discuss the implications of oversampling young ages in each age-group for rubella serological surveys.

The magnitude of effect of a growing population demographic bias on ASFOI estimation was determined by age-group designations in the survey design and the shape of ASFOI. As stated above, in the fractional polynomials and penalized regression splines analytic methods, the force of infection (\(\dot{\lambda}(a)\)) was defined as the derivative of the proportion seropositive (\(\dot{\pi}'(a)\)) divided by the proportion seronegative (\((1 - \pi(a))\)). Figure 3.10 shows the relationship of the force of infection by age broken down into its defined numerator and denominator (see equation 3.8). The grey vertical lines in Figure 3.10a highlight the change in slope of \(\dot{\pi}'(a)\) from positive to negative. The peaks of \(\dot{\pi}'(a)\) are 7, 5, and 4 years old for each population with assumed \(R_0\) values of 3, 5, and 7, respectively. The effect of grouping ages older than these peak ages in a growing population resulted in overestimates of both the derivative of proportion seropositive (\(\dot{\pi}'(a)\)) and the proportion seronegative (\((1 - \pi(a))\)) (Figure 3.10). The resultant effect on the ASFOI was less straightforward, an overestimate in \(\dot{\pi}'(a)\) biased the ASFOI upwards, and an overestimate in \((1 - \pi(a))\) biased the ASFOI downwards. We found in our three populations with varying values of \(R_0\), the overall effect on ASFOI estimates depended on the age-groupings and analytic method. Understanding ASFOI estimate biases help us to dissect estimate bias of \(R_0\) and the CRS incidence rate. We summarize our findings below by referencing Figure 3.10 to explain direction of biases; for detailed results of survey defined age-group biases, see Appendix C.1.
Figure 3.10: Estimation of ASFOI from proportion seropositive broken down into defined numerator and denominator.  a) The derivative of the proportion seropositive by age, $\hat{\pi}'(a)$ (numerator).  b) Proportion seronegative by age, $(1 - \hat{\pi}(a))$ (denominator).

We first looked at four age-grouping schemes that grouped ages 1 to 10 years (see Appendix C.1.1). When grouping ages 1 to 10 years, the peaks in the derivative of the proportion seropositive by age, as seen in Figure 3.10a, played an important role in understanding the effect of age-group bias on estimates of ASFOI when using penalized regression splines and fractional polynomials. In a growing population, grouping ages that are less than these peaks, e.g., a 1-5 year age-group when $R_0$ is 5, resulted in an underestimate of the derivative of proportion seropositive ($\hat{\pi}'(a)$), and an overestimate in the proportion seronegative ($((1 - \hat{\pi}(a)))$) (Figure 3.10). Taken together, these two biases resulted in a large underestimate of ASFOI for this age-group. Grouping ages that are older than the peak age of $\hat{\pi}'(a)$, means that $\hat{\pi}'(a)$ was overestimated, and ASFOI also tended to be overestimated. When using penalized regression splines, these two biases counteracted one another and the estimate of $R_0$ and the CRS incident rate was similar to the best analytic estimate. Using the less flexible parametric fractional polynomial approach, the magnitude of underestimation for the 1-5 year age-group drove the ASFOI biases for the other ages upwards, resulting in higher $R_0$ biased towards the true parameter values. The CRS incidence rate estimates across age-groupings between 1 and 10 years old was inconsistent across different $R_0$ simulations. Using the piecewise constant analytic method, as compared to the best analytic estimates above, grouping ages 1-5 and 5-10 caused the younger age FOI ($\hat{\lambda}_y$) to be overestimated, resulting in underestimates of the older age FOI ($\hat{\lambda}_o$). As a result, $R_0$ estimates increased toward the true parameter value, and CRS
incidence rates estimates decreased toward the true parameter value.

We concluded that when using fractional polynomials and piecewise constants methods, five year age-groupings between 1 and 10 years old resulted in mis-estimates migrating away from the best analytic estimates, but generally towards the true parameter values (with the exception of the CRS incident rate estimate for fractional polynomials when $R_0$ was 7). Given the accuracy of generalized regression splines under perfect data, age-groupings between ages 1 and 10 years old resulted in biases away from both the true parameter estimates and the best analytic estimates. However, estimates of the ASFOI, CRS incidence rate, and $R_0$ were still more accurate using penalized regression splines than using fractional polynomials or piecewise constant methods. As a result, in our final analysis we will group ages 1 to 10 years old into popularly used five year age-groups.

We secondly explored four age-grouping schemes that grouped ages 10 to 20 years, and found agreement in the direction of bias for all analytic methods and assumed population $R_0$ values (see Appendix C.1.2). For 10 to 20 year age-groupings, the overestimate in the proportion seronegative $((1 - \pi(a)))$ in a growing population mostly drove the bias in the ASFOI estimates downwards using both the fractional polynomial and penalized regression splines analytic methods (Figure 3.10b). The same was true when using the piecewise constant analytic method; underestimation of proportion seropositive caused the younger age FOI ($\bar{\lambda}_y$) to be underestimated. As a result, as age-groups widened between the ages of 10 and 20, estimates of $R_0$ decreased, moving further away from both the best analytic estimate and true parameter value. Additionally, as age-groups widened between 10 and 20 years old, estimates of the CRS incidence rate increased away from the best analytic estimate due to a larger proportion of susceptible individuals at early reproductive ages. In the piecewise constant method, the increase in the CRS incidence rate was also due to overestimates in the older age force of infection ($\bar{\lambda}_o$) to compensate for underestimates in lower age force of infection ($\bar{\lambda}_y$).
We concluded that age-groupings between ages 10 and 20 biased parameter estimates for all analytic methods away from true parameter values. It was important to include at least two age-groups between 10 and 20 years old to reduce age-group bias, and the more age-groups included between 10 and 20, the less biased were the estimates of epidemiological parameters. The literature shows that 13 years old may theoretically be a good choice for a break point, because this is when children stop attending school in many countries and age contact patterns change (Vynnycky et al., 2016). However, for rubella, 15 years old is also an important break point to distinguish reproductive ages from non-reproductive ages. We explored two age-groupings for ages 10 to 20 in the final analysis: (10,13,15,20), and (10,15,20).

Third, we looked at four age-grouping schemes that grouped ages 20 to 50 years (see Appendix C.1.3). For fractional polynomials, the overestimate in \( \hat{\pi}'(a) \) at late ages drove the bias in the ASFOI (Figure 3.10a). The direction of bias was the same using the piecewise constant approach; as age-groups widened between ages 20 and 50, the estimates for FOI at older ages (\( \bar{\lambda}_o \)) increased given the greater influence of the data for ages 13 to 20 years old. Subsequently, \( R_0 \) estimates moved upwards towards the true parameter values for both the fractional polynomials and piecewise constant methods. An increase in the older age ASFOI resulted in a lower proportion seronegative, i.e., fewer individuals susceptible to infection in reproductive age. The overall effect on the estimated population CRS incident rate was dependent on the how quickly the susceptible population depleted. At a higher transmission rate (assumed \( R_0 \) of 7), the susceptible individuals depleted faster and the CRS incidence rate was underestimated toward the true parameter value. At a lower transmission rate (assumed \( R_0 \) of 3), the susceptible individuals depleted more slowly and the population CRS incident rate was overestimated. For the penalized regression splines model, the overestimate in \( (1 - \hat{\pi}(a)) \) at late ages drove the bias in the ASFOI downwards, resulting in an underestimate of \( R_0 \) away from truth (however, the difference from the truth was very small, ranging from 0.03 to 0.06). We found that grouping ages 20 to 50 years into one large age-group underestimated the CRS incidence rate when using penalized regression splines. However, allowing a 20 to 25 year age-group reduced this bias because a data break here allowed for the detection of
the second peak in the true ASFOI, pulling late age ASFOI back up.

We concluded that fractional polynomials and piecewise constant analytic methods permitted grouping of ages 20 to 50 years old into one large age-group, because biases tended to move away from the best analytic estimates towards the true parameter values. Using penalized regression splines, age-groupings between ages 20 and 50 years old resulted in biases away from true parameter estimates and best analytic estimates. If at least one five year age-group (20 to 25 years old) was included, then the size of the CRS estimation bias diminished. Therefore, we explored two age-groupings for ages 20 to 50 years in the final analysis: (20,50], and (20,25,50).

Based on our age-group bias results, we selected four final age-grouping schemes to both minimize the number of age-groups and age-group bias. Selected groups included: $AG_{13}: [1, 5, 10, 13, 15, 20, 25, 50]$, $AG_{14}: [1, 5, 10, 13, 15, 20, 50]$, $AG_{15}: [1, 5, 10, 15, 20, 25, 50]$, and $AG_{16}: [1, 5, 10, 15, 20, 50]$. It is important to note that all final age-grouping schemes biased epidemiological parameter estimates away from the best analytic estimates described above; the overall direction and magnitude of bias depended on the analytic method used and the true age immunity profile of the population. Appendix C.1.5 shows results using perfect seroprevalence data for each age-group, i.e., the average proportion seropositive per defined age-group assigned to the mid-age of the age-group. We discuss biases of the final four age-grouping schemes using simulated serological surveys in the following section.

Focusing on surveys that only sample from women of childbearing age, we last examined five age-grouping schemes among reproductive age serological data. We found that by comparing parameter estimates to best analytic estimates, the added bias from age-grouping was minimal for all populations and analytic methods, as long as at least three age-groups spanning ages 15 to 50 were included. However, biases became more significant when as few as two age-groups were used, especially when the assumed $R_0$ was equal to 3 (see Appendix C.1.4).
3.4.3 Sampling Bias

Figures 3.11, 3.12, and 3.13 show the results for epidemiological parameters estimated from simulated serological data using the fractional polynomials analytic method for the final four age-grouping schemes for three populations with assumed $R_0$ values of 3, 5, and 7, respectively. Using the fractional polynomials analytic method, including the 10 to 13 age-group and not including the 20 to 25 year age-group (AG14), was preferred as it generally resulted in a more accurate mean ASFOI curve as compared to the best estimate from the fractional polynomial method. The resulting direction of estimation bias for $R_0$ and the CRS incidence rate was away from the best analytic estimates and toward the true parameter values. However, the variance of parameter estimates was larger when the age-group 20 to 25 was not included (AG14 vs. AG13), likely due to larger sampling variance across a wider age-group. At lower values of assumed population $R_0$, smaller samples sizes tended to bias mean estimates of CRS incidence upwards and $R_0$ downwards, but when $R_0$ equals 7 there was little difference observed across sample size. Age grouping 14 (AG14) is recommended for reducing parameter estimation biases using fractional polynomials, despite its slightly larger variance as compared to AG13.

Figures 3.14, 3.15, and 3.16 show the results of epidemiological parameter estimates using the two-age-group piecewise constant analytic method for the final four age-grouping schemes for three populations with assumed $R_0$ values of 3, 5, and 7, respectively. Across these final four age-grouping schemes, we found that the addition of more age-groups provided better parameter estimates when analyzing seroprevalence data with the two-age-group piecewise constant method, although including a 13 year age break reduced parameter estimation bias more than including the 25 year age break. Similar to the fractional polynomials estimates, piecewise constant estimates had greater uncertainty in older age FOI ($\hat{\lambda}_o$), for age-groupings that did not include the 25 year age break compared to those that did (AG14 vs. AG13 and AG16 vs. AG15). We also found more uncertainty in estimates of younger and older age FOI ($\hat{\lambda}_y$ and $\hat{\lambda}_o$) as the assumed population $R_0$ increased. Age grouping 13 (AG13) is recommended in order to reduce both parameter estimation bias and variance using piecewise constant approach.
Figure 3.11: $R_0 = 3$, Epidemiological parameter estimates based on simulated serological data using fractional polynomials. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figure 3.12: $R_0 = 5$, Epidemiological parameter estimates based on simulated serological data using fractional polynomials. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figure 3.13: $R_0 = 7$, Epidemiological parameter estimates based on simulated serological data using fractional polynomials. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figure 3.14: $R_0 = 3$, Epidemiological parameter estimates based on simulated serological data using two-age-group piecewise constants. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figure 3.15: $R_0 = 5$, Epidemiological parameter estimates based on simulated serological data using two-age-group piecewise constants. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figure 3.16: $R_0 = 7$, Epidemiological parameter estimates based on simulated serological data using two-age-group piecewise constants. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figures 3.17, 3.18, and 3.19 show the results of epidemiological parameter estimates using penalized regression splines for the final four age-grouping schemes for three populations with assumed $R_0$ values of 3, 5, and 7, respectively. As found above in the age-group bias analysis, it was important to break up the 20 to 50 year age-group with a 20 to 25 year age-group in order to reduce estimation bias of the CRS incidence rate, but the effect of not including the 20 to 25 age group had minimal effect on $R_0$ estimates (AG13 vs. AG14 and AG15 vs. AG16). Including the 10 to 13 year age-group resulted in more accurate mean estimates of $R_0$ (AG13 vs. AG15 and AG14 vs. AG16).

Similar to the piecewise constant analytic method, we saw larger uncertainty in estimation of the ASFOI and $R_0$ as the assumed population $R_0$ increases, specifically for the smallest sample size (60 samples per age-group). For example, the upper bound of the 95%CI of $R_0$ for AG16 was 74 (not shown in Figure C.45f). Age groupings 13 (AG13) or AG15 are recommended for reducing both inference bias and variance when using penalized regression splines.

3.5 Discussion

We explored in detail the sources of bias for serological surveys to support efforts in CRS burden estimation and vaccine program design. Our choice of using a growing population and our evaluation of a range of assumed $R_0$ values, resulting in a range of ASFOI, meaning that specific recommendations discussed here are likely to be applicable in many countries, where: i) past serological surveys are being analyzed, or ii) serological surveys may occur prior to the introduction of rubella-containing vaccine. The following sections summarize our findings for each type of bias, then we discuss limitations of our analysis, and make suggestions to direct future work.

Analytic Method Bias: In our first analysis, we assessed analytic method bias alone by controlling for age-group and sampling bias. We found that the more flexible analytic method (i.e., penalized regression splines) most accurately estimated rubella epidemiological parameters from both all-age and reproductive-age serological data as compared to the more popular analytic methods of fractional polynomials and two-age-group piecewise constants. Due to constraints of the model forms, both the fractional polynomials and two-age-group piecewise constant methods were unable to
Figure 3.17: $R_0 = 3$, Epidemiological parameter estimates based on simulated serological data using penalized regression splines. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figure 3.18: $R_0 = 5$, Epidemiological parameter estimates based on simulated serological data using penalized regression splines. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
Figure 3.19: $R_0 = 7$, Epidemiological parameter estimates based on simulated serological data using penalized regression splines. Figures a-d show the proportion seropositive and the ASFOI for the final 4 Age Groupings (AG): the grey vertical lines represent the age-groups, the red lines represent the true proportion seropositive and the ASFOI, dashed grey lines and light grey shaded area represent the mean estimated proportion seropositive and the ASFOI and 95% confidence interval from simulated serological surveys with sample size of 60 per age-group, dashed green lines and light green shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 250 per age-group, and dashed blue lines and light blue shaded area represent the mean estimated proportion seropositive and ASFOI and 95% confidence interval from simulated serological surveys with sample size of 1000 per age-group. Figures e and f show dot plots of estimated CRS incidence per 100,000 live births and $R_0$, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate based on perfect seroprevalence data at every age in years.
capture the nuance in the proportion seropositive and the ASFOI necessary for accurate estimation of $R_0$ and CRS incidence rates. The fractional polynomials model overestimated CRS incidence rates and underestimated $R_0$ for both all-age or reproductive-age seroprevalence data because it generally underestimated the pre-reproductive age ASFOI and overestimated the ASFOI among the most fertile ages. In contrast, the two-age-group piecewise constant analytic method performed differently whether provided with all-age or reproductive-age seroprevalence data.

We chose to use the two-age-group (13,50] piecewise constant approach because it was used by Vynnycky et al. (2016) to make global estimates of CRS incidence for the WHO. Using perfect seroprevalence data in our models at all-ages in years, the piecewise constant approach overestimated the CRS incidence rate. The magnitude of the CRS incidence rate overestimation increased as the ASFOI curve shifted upwards (overestimated by 10, 27, 38% for population assumed $R_0$ values of 3, 5, and 7, respectively), and as the ASFOI curve shifted to the right. In direct contrast, providing the model with perfect seroprevalence data for reproductive ages resulted in underestimates of the CRS incidence rate by 5, 13, and 14%. Vynnycky et al. (2016) relied on both all-age and reproductive-age serological data to estimate the global CRS burden. While we show that the two-age-group piecewise constant approach certainly biases estimates of CRS incidence, the changing direction of the biases based on input data, combined with the overall sensitivity of this method to the ASFOI curve, makes it difficult to evaluate what the overall implications are for the current estimates of the global burden of CRS. However, it is possible to make country-specific interpretations of estimates. For example, in India one of the nine serological datasets only sampled reproductive age individuals, therefore the WHO estimates likely overestimate the CRS incidence rate for India. In contrast, CRS estimates from Benin, Congo, Ghana, Kenya, Madagascar, Mozambique, Nigeria, Senegal, and South Africa relied only on serological samples from reproductive-age women to estimate the CRS burden, and as a result the WHO estimates likely underestimate CRS incidence rates in these countries. It is important to note that, though not evaluated in this analysis, the general piecewise constant approach of estimating the force of infection can be highly flexible in estimating the FOI for many finely resolved age-groups because no explicit parametric form is assumed for each age-group. However, the many-
age-group piecewise constant approach requires more finely resolved age-group seroprevalence data.

Survey Defined Age Group Bias: Grouping any ages in a growing population resulted in inherent biases in parameter estimates away from the best analytic estimate (based on the most finely resolved seroprevalence data by age in years), because younger ages per defined age-group were over-represented. Given the accuracy of estimates using penalized regression splines, biases from grouping ages moved in the direction away from both the best analytic estimates and the true parameter values. Therefore, finding the simplest survey design (fewest age-groups and minimal bias), relied upon minimizing biases overall. In contrast, the best analytic estimates using fractional polynomials and piecewise constant methods were inherently biased. Finding the simplest survey design meant prioritizing age-group biases that moved closer to the true parameter values over age-group biases that moved estimates further away from both the true and best analytic estimate for fractional polynomial and piecewise constant methods. We found that the popularly used five year age-groups between 1 and 20 years, plus a 30 year age-group to 50 was powerful for estimating $R_0$ and the CRS incidence rate (AG16), however adding an age break at 13 and/or 25 years (depending on the analytic method) pulled biased estimates back towards the truth with minimal sampling effort. This conclusion must be applied cautiously when using fractional polynomials and two-age-group piecewise constant methods because we are ultimately relying on age-group bias to correct for analytic bias. As such, it is very important to understand the underlying mechanisms of bias, as laid out in this analysis. We recommend careful determination of serological survey designs and evaluation of estimates of past and future serological survey results.

Sampling Bias: Using simulation tools to create ‘realistic’ serological data by sampling from the population at random, we estimated parameters for three age-group sample sizes (1000, 250, and 60) based on estimate precisions of 95, 90, and 80%, respectively. In addition to the expected finding that the variance of estimates increased as sample size decreased, we also found that variance of estimates increased across all sample sizes as rubella transmission in the population increased. At higher rubella $R_0$ values, proportion seropositive increased more steeply leaving less
information for the model to estimate the proportion seropositive, and the ASFOI. Therefore, if the 
$R_0$ of rubella is suspected to be higher than the median in Africa (5.2), then a smaller estimation 
precision and larger sample sizes are recommended. Despite the overall better performance of 
penalized regression splines as compared to fractional polynomials and two-age-group piecewise 
constant methods, we found larger uncertainty in estimates using penalized regression splines. 
As a result, it may be insightful to fit multiple analytic models to serological data for comparability.

Limitations: The three ASFOI curves we simulated should be broadly representative of rubella 
dynamics because their shapes were based on a widely-used age contact matrix estimated from the 
European POLYMOD diary study (Mossong et al., 2008), and the overall level of transmission was 
based on an empirically estimated rubella $R_0$ value (Lessler and Metcalf, 2013). The high rates of 
contact among assortative ages and between parent and child ages identified by the POLYMOD 
study and used in this analysis have also been echoed in other settings such as China and Vietnam 
(Read et al., 2014; Horby et al., 2011). Our assumed $R_0$ values of 3, 5, and 7, were based on 
African continent wide-estimates of rubella $R_0$ (median: 5.12, 90% CI 4.0-6.7) (Lessler and Metcalf, 
2013). The result was three different age immunity profiles whose average ages of infection ranged 
between 5 and 10 years old, capturing empirical estimates (Anderson and May, 1991; Grenfell 
and Anderson, 1985). Therefore, our conclusions above can be roughly generalizable to rubella 
serological surveys. However, region-specific ASFOI curves may still remain outside our explored 
range; for example, a higher rubella $R_0$ ranging between 6.9 and 11.8 has been estimated in Addis 
Ababa, Ethiopia (Cutts et al., 2000). If the overall level of transmission is higher than in the 
population explored here, and our assumed age contact matrix is correct, we would generally expect 
biases to follow trends across $R_0$ revealed in our analysis, i.e., with increases in the magnitude 
of mis-estimation of epidemiological parameters. Additionally, larger sample sizes per age-group 
will be needed to increase estimate precision. The effect of a significantly different region-specific 
age contact matrix than explored here is more difficult to evaluate. As we learn more about age 
contact patterns, further exploration of inference bias from serological surveys will be valuable.
The same population demography was used for all analyses, that is a stable population growing with a rate of 1.7% per year. We chose a growing population because all countries that currently have not introduced the rubella vaccine and have endemic rubella are growing populations. Vynnycky et al. (2016) estimated that the largest burden of CRS cases is in India, motivating our decision to use population demographic parameters based on the Indian state of Bihar for our simulated populations. However, other countries who have not introduced the rubella vaccine have higher growth rates upwards of 3.5% mean growth per year (United Nations and Department of Economic and Social Affairs and Population Division, 2015; World Health Organization, 2016c). Larger population growth rates decrease the average age of infection and increase the equilibrium of the proportion of infected individuals (Tuljapurkar and John, 1991), thereby increasing the ASFOI. As stated above, if the ASFOI curve is larger than explored here, assuming the shape of the curve or the age contact matrix is correct, then the magnitude of biases will follow demonstrated trends above, and will likely increase, and larger sample sizes are recommended.

Previous work suggests that we can ignore seasonality when estimating the ASFOI from serological data in stable populations (Whitaker and Farrington, 2004). However, Ferrari et al. (2010) showed in Niger that high birth rates and erratic outbreaks can break the assumption of ASFOI time homogeneity within the context of seasonal changes for measles. A changing age distribution in demographically non-stable populations can also break the assumption of a time invariant ASFOI (McLean and Anderson, 1988). For example, Manfredi and Williams (2004) modeled measles dynamics in Italy allowing age transmission rates to be additionally scaled by the changing age distribution of the population, and found significant differences in measles dynamics and necessary vaccination coverage. Ideally, with serological surveys that are collected over a broad time span, the ASFOI can be estimated as time varying in order to reduce estimation bias, and we would generally expect biases similar to what we have demonstrated in this analysis. However, for cross-sectional serological surveys the timing of the survey will affect the estimation of ASFOI. Exploring the best timing of a cross-sectional serological survey given a time changing ASFOI is an important research question for future work. Albeit beyond the scope of this analysis, it is
likely to be time and region specific.

For simplicity, we ignored the effects of waning rubella immunity (Reef and Plotkin 2013) and error in diagnostic testing by assuming perfect knowledge of seropositivity based on past infection history. We also ignored maternal immunity by dropping data of individuals under the age of 1 years old, because the inherited antibodies result in a high proportion seropositive and a low hazard of infection. Extending this analysis to account for waning and maternal immunity is recommended for future work. However, we anticipate that the broad scope of bias and the underlying mechanistic drivers as revealed here, will still hold.

Biases are likely in parameter estimates from serological data in complex infectious disease dynamics. Understanding these biases is an important endeavor for researchers and public health officials who are making policy recommendations based on analyses of serological data. This paper explored the roles that population dynamics, analytic method, and survey design (i.e., age-groupings and sample size) play in biasing rubella epidemiological parameters of interest via simulation tools. We detailed the nuanced relationships between these characteristics and bias that can be used by public health officials to understand past serological survey estimates. We additionally recommended that rubella serological survey designs be used to evaluate current CRS burden and $R_0$ in order to aid in decision for introduction of rubella vaccine. Given the detrimental consequences of CRS, and the Global Alliance for Vaccines and Immunizations (GAVI) pledged support of rubella vaccine introduction (Burki 2012; Global Alliance for Vaccines and Immunizations 2014), this is a timely investigation as eligible countries consider rubella vaccine introduction.
The introduction of the measles and rubella vaccines were the sole catalyst for advancement towards control and elimination of measles, rubella, and Congenital Rubella Syndrome (CRS). For example, the World Heath Organization estimated in 2015 that measles vaccination saved 17.1 million lives, and, measles and rubella elimination has been achieved in the WHO region of the Americas since 2002 and 2009, respectively (CDC 2015, 2014). However, measles remains a significant cause of morbidity and mortality in the world; twenty million individuals are infected with measles annually and an estimated 115,000 measles deaths occurred globally in 2014 (World Health Organization 2016a). Additionally, the devastating consequences of rubella infection among pregnant women continue to affect families around the world; Vynnycky et al. (2016) estimated 105,000 CRS cases occur annually. Successful programs require continued vigilance, and as the epidemiological context changes, the power of different data sources will also shift. Mathematical models allow us to anticipate and understand these epidemiological shifts and incorporate new data sources, and thus play an important role in the continued decision-making process for the measles and rubella control effort.

This dissertation explored public health questions related to the current status of measles and rubella incidence and immunity, and the control of measles and rubella. While each chapter is intended to be a separate publishable manuscript, three common themes are found throughout. Each chapter discussed direct policy implications and recommendations for national vaccine control
programs. In chapter 1, a paper originally motivated by a request from the National Technical Advisory Group on Immunizations in India, we identified high CRS burden regions in India and listed specific risk factors that can be used to determine regions at risk of increases in CRS burden post RCV introduction. In chapter 2, we used a new data source to estimate measles susceptibility by age in Madagascar, and then recommended specific age targets for SIAs in order to prevent the potentially looming measles outbreak. In chapter 3, we simulated rubella endemic populations and serological surveys, and used these simulations to estimate rubella epidemiological parameters. We discussed in detail how to better interpret estimates from past serological surveys, and recommended a simple survey design for implementing future serological surveys when considering introduction of rubella-containing vaccine (RCV).

The second common theme across chapters was a focus on population demography. A plethora of literature has shown that demography, specifically the population age structure and rate of births for childhood infectious diseases, directly and significantly affect transmission dynamics (John, 1990b,a; Tuljapurkar and John, 1991; Manfredi and Williams, 2004; Anderson and May, 1991). Inherent non-linearities mean that the relationship between population demography and infectious disease transmission is often complicated. For example, lower birth rates result in higher average age of infection of rubella and therefore a higher CRS burden. If the rubella vaccine is introduced in this environment, the critical threshold of vaccination coverage (i.e., necessary coverage to prevent increases in CRS) is lower because it requires less vaccination to reduce transmission of rubella, and because this results in a marked reduction of CRS cases compared to a high starting burden (Metcalf et al., 2012a). Despite the empirically and theoretically proven importance of demography in infectious disease dynamics, many modelers continue to rely on simplistic demographic assumptions, such as assuming constant birth rates over time, or stable population dynamics. In chapters 1 and 2, we matched our simulated populations in India and Madagascar as closely as possible to known demographic features by extracting probabilistic projections of birth rates based on the UNPD 2015 revisions (United Nations, 2015). In chapter 3, even though we assumed stable population dynamics to control for these changes on estimation
bias, we discussed how non-stable populations would affect the results.

Third, all chapters included a focus on serological data (i.e., the need for serological data, the use of serological data, and the how to interpret and collect serological data, across chapters 1-3, respectively). Serological data, specifically for immunizing infections, allows inference of key epidemiological parameters that inform transmission mechanisms in mathematical models. Chapter 1 called for the need of population-based serological surveys in India to estimate the basic reproductive number of rubella and risk heterogeneity, in order to increase regional accuracy of predictions on the consequences of rubella vaccine introduction. Chapter 2 explored new ways to use existing serum samples for serological surveys. Although we were unable to recommend nested serological surveys within measles and rubella surveillance systems for general use, this analysis took the first step by comparing direct estimates to indirect estimates from age cohort immunity projections. Last, chapter 3, discussed specific strengths and limitations of rubella serological surveys to estimate important rubella epidemiological parameters.

This dissertation adds to the existing body of literature that uses mechanistic mathematical models to explore the current status of measles and rubella incidence and immunity, and the control of measles and rubella. Throughout the dissertation, we offered specific recommendations for future work related to the topic of each individual chapter. The following paragraphs recall and expand upon these directions for future research.

Future research directions for the control of rubella and prevention of CRS

I will first discuss avenues of future research related to estimates of CRS incidence and control of rubella, and prevention of CRS. As of 2015, the rubella vaccine had been introduced into the national immunization schedules in 149 of 194 WHO member states ([World Health Organization, 2015](https://www.who.int/immunization/monitoring_surveillance/diseases/rubella/en/)). Due to the estimated high burden of CRS by [Vynnycky et al. (2016)](https://www.ncbi.nlm.nih.gov/pubmed/27252405), and the Global
Alliance for Vaccines and Immunizations’ support of the introduction of the RCV, many countries are currently planning or considering introducing the rubella vaccine (Burki, 2012) (see Figure 3.20). However, it is possible that the timing of this international push to introduce RCV into childhood immunization schedules is inopportune for some countries. Of the remaining 45 countries who have not introduced RCV, 41 are African countries, and 4 are Asian countries. Comparing Figure 3.20 to Figure 3.21 it is clear that the majority of countries that have not yet introduced RCV have some of the highest global birth rates (with the exception of India and Indonesia). Mathematical models of rubella show that higher birth rates result in lower average ages of infection for rubella, smaller CRS burden, and a higher risk of short-term CRS increases post rubella vaccine introduction into childhood immunization schedules (Metcalf et al., 2012a). Therefore, the best timing of RCV introduction into childhood immunizations schedules may be many years from now in countries when declining birth rates begin to stabilize. Estimating the CRS burden beyond those produced by Vynnycky et al. (2016) and modeling the effect of introducing the rubella vaccine in these specific countries is a timely direction for future research.

A core feature which motivated much of the work in chapter 3 is that, currently, global estimates of CRS rely on model extrapolations from serological surveys. Novel ways to estimate CRS burden would provide a powerful improvement in our understanding of this pathogen. We suggest the following approaches: 1) Incorporate findings from chapter 3 regarding best analytic method and the direction and magnitude of biases of CRS estimates from serological data to analyze and interpret results using existing IgG serological surveys. 2) Explore new avenues for IgG serological samples from existing serum, such as nested serological samples within HIV surveillance systems. This type of data may be particularly be useful in countries where the HIV is endemic in the general population. 3) Explore estimation of appropriate denominators to case-notification CRS cases by using small controlled surveillance sites.

A number of extensions to modeling the introduction of RCV and vaccine delivery systems are also suggested by this work: 1) Incorporate probabilisitic projections of changing fertility over
Figure 3.20: Countries with Rubella vaccine in the national immunization programme; and planned introductions 2016-2017. (World Health Organization, 2015)

Figure 3.21: Map of births per 1,000 people (United Nations, 2015).
time into mathematical models to estimate best timing of RCV introduction. For example, even though Somalia’s total fertility rate is slightly higher than Mali’s as of 2015 (6.7 vs. 6.3), Somalia’s fertility decline is predicted to occur at a faster pace. The rate of fertility decline will affect the slope of rise in CRS incidence, and will play a role in determining the optimal timing of rubella vaccine introduction. 2) Explore the option of vaccinating women of childbearing age for rubella. In the WHO 2011 rubella position paper, the WHO encourage countries that aim to interrupt rubella transmission with childhood immunization programs to also administer vaccination campaigns that target women of reproductive age so as to avoid the risk of potential increases in CRS\footnote{World Health Organization\citeyear{WorldHealthOrganization2011a}.} However, many GAVI eligible countries, including India, are not considering this recommendation because women of childbearing age are a difficult population to reach, and because GAVI’s support does not offer subsidies to vaccinate this age group of reproductive age women\footnote{Burki\citeyear{Burki2012}.} Research that evaluates i) avenues for reaching women of reproductive age (i.e., postpartum visits or child measles vaccination visits), in combination with ii) estimates of CRS risk by age using serological data, can evaluate the potential of vaccinating this sub-group to reduce the risk of CRS increases. This direction of research may be particularly important for countries planning to introduce the vaccine into childhood immunization schedules but also have high birth rates, i.e., some regions of India as discussed in chapter 1, Zambia, or Kenya. 3) Conduct case studies in Costa Rica and Italy by comparing vaccination rates of rubella before and after short term increases in CRS incidence, in order to evaluate potential changes in public opinion toward rubella-containing vaccine. Current mathematical models only consider the immediate effect of increases in CRS incidence, however, there may be longer-term effects on vaccine immunization uptake if public trust in the rubella vaccine diminishes.

\textbf{Future research directions for social sciences in infectious disease modeling}

The second broad future direction of research extends the role of demography and social sciences in infectious disease modeling. Ultimately, this is an exciting time for social scientists in the field of infectious disease modeling. For example, big data is being used to estimate the spatial contact
patterns vital to the evaluation of heterogenous disease transmission (Wesolowski et al., 2015). An additional extension of infectious disease modeling, as mentioned above, is incorporating probabilistic projections of demographic parameters. The UN world population projections no longer rely upon sensitivity analysis of demographic parameter changes based on expert opinion, instead the 2015 projections were revised by using a statistical Bayesian hierarchical model to estimate demographic parameters (United Nations, 2015; United Nations and Department of Economic and Social Affairs and Population Division, 2015). The R packages ‘bayesPop’, ‘bayesFR’ and ‘bayesLife’, created by Hana Sevcikova and Adrian Raftery and team at the University of Washington, make incorporating probabilistic projections into infectious disease models easily accessible (Raftery et al.).

Future research directions for serological survey data collection and analyses

The last broad direction of future research is related to the use and potential of IgG serological data. Serological surveys are a powerful resource for understanding infectious disease dynamics. For immunizing infections like rubella and measles, this data source provides a window into the landscape of population immunity. In countries where vaccines are administered, serological surveys can be used to evaluate vaccination programs, determine if supplementary immunization activities are necessary and if so, the age-group to be targeted, and can even predict future outbreaks (Andrews et al., 2008; Babad et al., 1995). Serological surveys can additionally inform country-specific knowledge gaps of disease burden and unknown transmission dynamics (Ximenes et al., 2014; Chakravarty et al., 1976). In fact, Metcalf et al. (2016) proposed a World Serology Bank to drive a step change in our understanding and optimum control of infectious diseases. Given the importance of serological data, in chapter 3 we explored a cost-effective serum source for IgG serological testing. We concluded that an important direction for future research is replication of the estimated age immunity profile of nested measles IgG serological surveys within the measles surveillance system via population or community-based serological data. It will also be important to establish that this data source has comparable generalizability if i) the rubella vaccine is administered in the country, potentially changing the sample of patients presenting for care with fever and rash,
and ii) conducted in other country contexts with different measles and rubella vaccination coverage.

Measles and rubella remain significant causes of mortality and morbidity in the world. As the epidemiological and demographic context continues to change, mathematical models that can anticipate and understand epidemiological and demographic shifts will remain important decision making tools for the rubella and measles control effort.
Bibliography


British Broadcasting Company. India to provide four free vaccines, including rotavirus, 2014.


Hindustan Times. India adds 4 new vaccines to immunisation programme, 2014.


Appendices
Appendix A

List of Abbreviations

AIC       Akaike Information Criterion
AG        serological survey defined Age Groupings
ASFOI     Age-Specific Force of Infection
ASFR      Age-Specific Fertility Rates
BIC       Bayesian Information Criterion
CBR       Crude Birth Rate (births per 1000)
CRS       Congenital Rubella Syndrome
FP        Fractional Polynomials
GAVI      Global Alliance for Vaccines and Immunizations
IgG       Immunoglobulin G
IgM       Immunoglobulin M
IPM       Institut Pasteur de Madagascar
lb        lower bound of 95% confidence interval
MCV1      Measles-Containing Vaccine dose 1
mo        months old
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>bivalent Measles-Rubella Vaccine</td>
</tr>
<tr>
<td>MMR</td>
<td>trivalent Measles-Mumps-Rubella Vaccine</td>
</tr>
<tr>
<td>PRS</td>
<td>Penalized Regression Splines</td>
</tr>
<tr>
<td>PC</td>
<td>Piecewise Constant</td>
</tr>
<tr>
<td>RCV</td>
<td>Rubella-Containing Vaccine</td>
</tr>
<tr>
<td>ReproAG</td>
<td>serological survey defined Reproductive Age Groupings</td>
</tr>
<tr>
<td>SIA</td>
<td>Supplementary Immunization Activities</td>
</tr>
<tr>
<td>ub</td>
<td>upper bound of 95% confidence interval</td>
</tr>
<tr>
<td>UIP</td>
<td>India’s Universal Immunization Programme</td>
</tr>
<tr>
<td>UNICEF</td>
<td>United National Children’s Fund</td>
</tr>
<tr>
<td>UNPD</td>
<td>United Nations Population Division</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>yo</td>
<td>years old</td>
</tr>
</tbody>
</table>
Chapter 2 Supplementary Material

B.1 Supplementary Methods

B.1.1 Model Selection to Estimate Population Proportion Seropositive by Year

The objective was to estimate the total and under five years old proportion seropositive for each complete year in the data (2005-2015). To estimate the total and under five years old proportion seropositive by year, we first smoothed the data over age and then extracted age weighted average proportions seropositive based on Madagascar’s population age structure (United Nations, 2015). We modeled the proportion of individuals seropositive based on 4394 serum samples by age using a semiparametric approach. Semiparametric approaches are ideal for assessing seroprevelance by age in non-endemic settings and using non-probability sample because they are flexible and can ignore constraints such as monotonicity of seropositivity across age required in parametric models (Hens et al., 2012). We focused on semiparametric penalized regression smoothers based on the ‘mgcv’ library in R (Wood, 2006, 2011; R Core Team, 2016). Penalized regression smooths over a scatterplot based on a defined spline basis and penalty that controls the ‘wiggliness’ by measuring tradeoffs between smoothness and fit to the data. We first determined the ‘best’ penalized
regression smoothers model based on the lowest AIC, BIC, and residual deviance. We compared penalized regression smoothers model with different parameter inputs (including age, year the sample was collected, and the interaction between the two covariates), different smoothing bases (cubic regression splines, and p-splines), and with different link functions (complementary log-log and log odds) (Wood, 2006). We used a tensor product smoother on the interaction term because it is recommended when there are large scale differences between the two covariates (Wood, 2006). The basis dimension simply designates an upper bound on the flexibility of the fit. The smoothing parameter, however, estimated by minimizing the generalized cross validation score, controls the actual degrees of freedom and defines the ‘wiggliness’ penalty in the least squares fitting (Wood, 2006). The choice of the basis dimension was large enough to have enough degrees of freedom to represent the underlying data (Wood, 2006).

Supplemental Table B.1A and B.1B show the results of these model fits (models #1-7). We chose model #4 as the ‘best’ performing model because it had the lowest residual deviance and lowest AIC, although not the lowest BIC. Model #4 included an isotropic smooth function on age and year with a cubic spline base, and a tensor product smooth on the interaction term between age and year with a cubic spline base. After selecting model #4 as ‘best’ semiparametric model, we then compared it to parametric models (Supplemental Table B.1C). Given that the outcome variable is binary, we fit the data using a generalized linear model where the log odd on the proportion seropositive is a linear function of age. We also fit cubic B-splines on age (see ‘splines’ library in R) (R Core Team, 2016). Both models included a covariate for the year the serum was collected. The model fits of these three models are shown in Supplemental Table B.1C. The semiparametric model #4 continued to have the lowest AIC and residual deviance.

The ‘best’ performing model (model #4) is the most flexible model, and allows the age immunity profile to change by year, which we would expect given that population demography and rubella dynamics are changing overtime. Supplemental Figure B.1 shows the estimated proportion seropositive by age for years 2005 through 2015 based on model #4. These results highlight our reservations in the representativeness of the data because one would expect the curve of the age profile of im-
Table B.1: Overview of model fits. The top table A assesses the ‘best’ fit semiparametric model using penalized-regression smoothers, the second table B assesses the ‘best’ spline bases and link functions for the semiparametric model using penalized-regression smoothers, and the bottom table C assesses the overall ‘best’ model fit (Wood, 2006). N=4394

cloglog = complimentary log-log, logit=log odds

<table>
<thead>
<tr>
<th>A</th>
<th>model #</th>
<th>library</th>
<th>model</th>
<th>link</th>
<th>spline base</th>
<th>DF</th>
<th>BIC</th>
<th>AIC</th>
<th>Residual Deviance</th>
<th>chi-square test of RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>logit</td>
<td>cr</td>
<td>9.26</td>
<td>4023.85</td>
<td>3964.68</td>
<td>3948.2</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>R: mgcv</td>
<td>y ~ year + s(age)</td>
<td>logit</td>
<td>cr</td>
<td>10.3</td>
<td>3966.02</td>
<td>3900.8</td>
<td>3882.2</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>R: mgcv</td>
<td>y ~ s(year) + s(age)</td>
<td>logit</td>
<td>cr</td>
<td>17.18</td>
<td>3985.43</td>
<td>3974.71</td>
<td>3842.4</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>R: mgcv</td>
<td>y ~ s(year) + s(age) + ti(year*age)</td>
<td>logit</td>
<td>cr</td>
<td>43.99</td>
<td>4126.81</td>
<td>3845.78</td>
<td>3759.8</td>
<td>reference</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>model #</th>
<th>library</th>
<th>model</th>
<th>link</th>
<th>spline base</th>
<th>DF</th>
<th>BIC</th>
<th>AIC</th>
<th>Residual Deviance</th>
<th>chi-square test of RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>R: mgcv</td>
<td>y ~ s(year) + s(age) + ti(year*age)</td>
<td>logit</td>
<td>cr</td>
<td>43.99</td>
<td>4126.81</td>
<td>3845.78</td>
<td>3759.8</td>
<td>reference</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>R: mgcv</td>
<td>y ~ s(year) + s(age) + ti(year*age)</td>
<td>cloglog</td>
<td>cr</td>
<td>42.54</td>
<td>4118.44</td>
<td>3946.7</td>
<td>3763.6</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>R: mgcv</td>
<td>y ~ s(year) + s(age) + ti(year*age)</td>
<td>logit</td>
<td>ps</td>
<td>19.06</td>
<td>3976.89</td>
<td>3855.14</td>
<td>3819</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>R: mgcv</td>
<td>y ~ s(year) + s(age) + ti(year*age)</td>
<td>cloglog</td>
<td>ps</td>
<td>19.39</td>
<td>3982.67</td>
<td>3959.84</td>
<td>3822.1</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th>model #</th>
<th>library</th>
<th>model</th>
<th>link</th>
<th>spline base</th>
<th>DF</th>
<th>BIC</th>
<th>AIC</th>
<th>Residual Deviance</th>
<th>chi-square test of RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>R: base</td>
<td>y ~ year + age</td>
<td>logit</td>
<td>NA</td>
<td>13</td>
<td>4132.73</td>
<td>4049.68</td>
<td>4023.7</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>R: splines</td>
<td>y ~ year + bs(age)</td>
<td>logit</td>
<td>b-spline</td>
<td>15</td>
<td>4088.16</td>
<td>3972.34</td>
<td>3942.3</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>R: mgcv</td>
<td>y ~ s(year) + s(age) + ti(year*age)</td>
<td>logit</td>
<td>cr</td>
<td>43.99</td>
<td>4126.81</td>
<td>3845.78</td>
<td>3759.8</td>
<td>reference</td>
<td></td>
</tr>
</tbody>
</table>

Community to change across years in a wave-like fashion (i.e. a dip in immunity at age 5 years old in 2010 would result in a dip of equal size of less at age 6 years old in 2011, etc). Instead we see, for example, that the proportion seropositive at age 15 in 2011 is higher than the proportion seropositive at age 17 in 2013. These sorts of discrepancies in expected behavior of proportion seropositive are seen throughout the years by ages.

B.1.2 Model Selection to Estimate Current Proportion Seropositive by Age

The objective was to infer the current proportion of individuals seropositive by age. Despite the fact that model #4 estimates a distinct age profile of immunity curve for each year, including the most recent 2015, we decided not to rely on this model for two reasons: 1) Unexpected changes in estimated proportion seropositive by age across years. As Supplemental Figure B.1 shows, there are sometimes large discrepancies in expected wave-like behavior of proportion seropositive are seen throughout the years by ages. We know based on the total proportion seropositive analysis by year (see Figure 2.4), that the 2010-2015 data appears to better represent potential true changes based on known SIAs, likely due to larger sample sizes between 2010 and 2015. However, even
between 2014 and 2015, the changes in estimated proportion seropositive by age are not wave-like as expected (see Supplemental Figure B.1). 2) The sample size of 2015 data, specifically among 15 years old is considerably smaller than the previous two years. In 2015, 387 samples were collected; 34 samples were collected from individuals between ages 15 and 25 years old, and only 18 samples were collected from individuals between the ages 25 and 45 years old. As a result, we chose to combine all samples collected between November 2013 and December 2015 and then smooth over age without distinguishing between years to increase the sample size and potentially the generalizability of the sample. It was important not to combine samples across time when an SIA took place because it causes sharp shifts of seropositivity under the age of five years old. The result is a total sample size of 1083 samples, 92 samples were from individuals between 15 and 25 years old, and 72 samples were collected from individuals between ages 25 and 45 years old.

We again used a semiparametric penalized regression smoothers approach to model the current proportion of individuals seropositive by age because they are flexible and can ignore constraints
such as monotonicity of seropositivity across age required in parametric models (Wood, 2006; Hens et al., 2012). Given that we do not want to control for year, the simple model was already chosen to include a lone smoother on age. We selected the final model by comparing smoothing bases (cubic regression splines, p-splines, and thin plate regression splines), and link functions (complementary log-log and log odds) (Wood, 2006). The choice of the basis dimension, k, was large enough to have enough degrees of freedom to represent the underlying data (Wood, 2006).

Supplemental Table B.1A and B.2A show the results of these model fits (models #10-15). We choose model #10 as the ‘best’ performing model because it had the lowest residual deviance and lowest AIC, although not the lowest BIC. Model #10 included an isotopic smooth function on age with a cubic spline base and a logit link function. We then compared model #10 to parametric models (Supplemental Table B.2B). Given that the outcome variable is binary, we fit the data using a generalized linear model where the log odd on the proportion seropositive is a linear function of age. We also fit cubic B-splines on age (see ‘splines’ library in R) (R Core Team, 2016). The model fits of these three models are shown in Supplemental Table B.1B. The semiparametric model #10 had the lowest AIC, BIC, and residual deviance.

Table B.2: Overview of model fits for the estimation of current proportion measles seropositive by age. The top table A assesses the ‘best’ spline bases and link functions for the semiparametric model using penalized-regression smoothers, and the bottom table B assesses the overall ‘best’ model fit (Wood, 2006).  
cloglog = complimentary log-log, logit=log odds  
cr=cubic regression splines, ps = p-splines, tp=thin plate regression splines  
DF=degrees of freedom  
Chi-square test of RD=chi square test of difference in residual deviance (ref: model #10)  
***0.001, **0.01, *0.05, ns=non-significant

<table>
<thead>
<tr>
<th>A model #</th>
<th>library</th>
<th>model</th>
<th>link</th>
<th>spline base</th>
<th>DF</th>
<th>BIC</th>
<th>AIC</th>
<th>Residual Deviance</th>
<th>difference in RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>logit</td>
<td>cr</td>
<td>9.38</td>
<td>1110.96</td>
<td>1064.18</td>
<td>1047.40</td>
<td>reference</td>
</tr>
<tr>
<td>11</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>cloglog</td>
<td>cr</td>
<td>9.66</td>
<td>1112.59</td>
<td>1064.28</td>
<td>1046.90</td>
<td>ns</td>
</tr>
<tr>
<td>12</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>logit</td>
<td>ps</td>
<td>7.67</td>
<td>1103.38</td>
<td>1065.131</td>
<td>1051.8</td>
<td>ns</td>
</tr>
<tr>
<td>13</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>cloglog</td>
<td>ps</td>
<td>7.79</td>
<td>1104.74</td>
<td>1065.87</td>
<td>1052.3</td>
<td>ns</td>
</tr>
<tr>
<td>14</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>logit</td>
<td>tp</td>
<td>7.94</td>
<td>1106.53</td>
<td>1066.96</td>
<td>1053.10</td>
<td>*</td>
</tr>
<tr>
<td>15</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>cloglog</td>
<td>tp</td>
<td>7.97</td>
<td>1109.22</td>
<td>1069.48</td>
<td>1055.5</td>
<td>**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B model #</th>
<th>library</th>
<th>model</th>
<th>link</th>
<th>spline base</th>
<th>DF</th>
<th>BIC</th>
<th>AIC</th>
<th>Residual Deviance</th>
<th>difference in RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>R: base</td>
<td>y ~ age</td>
<td>logit</td>
<td>NA</td>
<td>2</td>
<td>1127.57</td>
<td>1117.59</td>
<td>1113.6</td>
<td>***</td>
</tr>
<tr>
<td>17</td>
<td>R: splines</td>
<td>y ~ bs(age)</td>
<td>logit</td>
<td>b-spline</td>
<td>4</td>
<td>1113.12</td>
<td>1093.16</td>
<td>1085.2</td>
<td>***</td>
</tr>
<tr>
<td>10</td>
<td>R: mgcv</td>
<td>y ~ s(age)</td>
<td>logit</td>
<td>cr</td>
<td>9.38</td>
<td>1110.96</td>
<td>1064.18</td>
<td>1047.40</td>
<td>reference</td>
</tr>
</tbody>
</table>
### B.1.3 Comparison and Direct and Indirect Empirical Analysis

Table B.3: Comparison between indirect and direct empirical estimates of proportion seropositive by age

<table>
<thead>
<tr>
<th></th>
<th>Indirect Estimates</th>
<th>Direct Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data source(s)</strong></td>
<td>vaccination coverage overtime (routine and SIAs)</td>
<td>measles IgG serological survey nested within measles/rubella surveillance system</td>
</tr>
<tr>
<td></td>
<td>maternal immunity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>vaccine efficacy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>attack rate overtime</td>
<td></td>
</tr>
<tr>
<td><strong>Estimation Technique</strong></td>
<td>combine birth cohort projection of immunity estimates based on each cohort’s experience of routine immunizations, SIAs, and natural infection</td>
<td>semiparametric approach using penalized regression smoothers</td>
</tr>
<tr>
<td><strong>Limitations</strong></td>
<td>relies of assumed parameters [Lessler et al., 2011a]</td>
<td>non-probability sample lacking external validity</td>
</tr>
<tr>
<td><strong>Strengths</strong></td>
<td>based on public access data</td>
<td>probabilistic estimates with standard errors, low cost extension to current surveillance system as compared to population-based serological survey</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td>[Takahashi et al., 2015]</td>
<td></td>
</tr>
</tbody>
</table>
B.2 Supplementary Results

B.2.1 Time Series of Measles Incidence across SIA Scenarios

The following figures show a time series of measles incidence by assumed value of $R_0$ (in rows) for all SIA scenarios. The black, red, and blue lines per figure represents three different starting age profiles of immunity based on serological data direct estimates; the mean represented by the black line, the modified 2.5% confidence lower bound (lb) represented by the red line, and the 97.5% confidence upper bound (ub) represented by the blue line. These time series of incident cases should not be considered predictions for measles outbreak sizes in Madagascar because we assumed a spatially homogenous mixing across the country of Madagascar. Various lines of evidence (including the presence of remote districts suggested by analysis of rubella incidence [Wesolowski et al., 2016], and the fact that measles cases have been occurring without any major outbreaks) suggest that this is inappropriate, and, consequently, the measles virus is likely to spread more evenly and quickly in the model than in reality. One approach to capturing the reduction in incidence as a result of heterogeneous spatial spread is to use a mapping between susceptibility and incidence based on Kalman Filtering based estimates of measles burden [Simons et al., 2012]. However, given that our focus is specifically on the nuance of the impact of age targeted SIAs, this simplification would obscure much of the variation of incidence. Therefore, we retain the homogenous framing and rely on order of magnitude to compare outbreak sizes between this no SIA scenario and SIA scenarios in the main document.
Figure B.2: Time series of measles incidence: no SIA
Figure B.3: Time series of measles incidence: SIA target children ages 9 months old to 5 years old and achieves 70% vaccination coverage.
Figure B.4: Time series of measles incidence: SIA target children ages 9 months old to 5 years old and achieves 80% vaccination coverage.
Figure B.5: Time series of measles incidence: SIA target children ages 9 months old to 5 years old and achieves 90% vaccination coverage
Figure B.6: Time series of measles incidence: SIA target children ages 9 months old to 5 years old and achieves 95% vaccination coverage.
Figure B.7: Time series of measles incidence: SIA target children ages 9 months old to 10 years old and achieves 70% vaccination coverage
Figure B.8: Time series of measles incidence: SIA target children ages 9 months old to 10 years old and achieves 80% vaccination coverage
Figure B.9: Time series of measles incidence: SIA target children ages 9 months old to 10 years old and achieves 90% vaccination coverage
Figure B.10: Time series of measles incidence: SIA target children ages 9 months old to 10 years old and achieves 95% vaccination coverage
Figure B.11: Time series of measles incidence: SIA target children ages 9 months old to 15 years old and achieves 70% vaccination coverage
Figure B.12: Time series of measles incidence: SIA target children ages 9 months old to 15 years old and achieves 80% vaccination coverage
Figure B.13: Time series of measles incidence: SIA target children ages 9 months old to 15 years old and achieves 90% vaccination coverage
Figure B.14: Time series of measles incidence: SIA target children ages 9 months old to 15 years old and achieves 95% vaccination coverage
Figure B.15: Time series of measles incidence: SIA target children ages 9 months old to 20 years old and achieves 70% vaccination coverage
Figure B.16: Time series of measles incidence: SIA target children ages 9 months old to 20 years old and achieves 80% vaccination coverage
Figure B.17: Time series of measles incidence: SIA target children ages 9 months old to 20 years old and achieves 90% vaccination coverage.
Figure B.18: Time series of measles incidence: SIA target children ages 9 months old to 20 years old and achieves 95% vaccination coverage
B.2.2 Time Series of $R_{eff}$ across SIA Scenarios

The following figures show a time series of $R_{eff}$ by assumed value of $R_0$ (in rows) for all SIA scenarios. The black, red, and blue lines per figure represents three different starting age profiles of immunity based on serological data direct estimates; the mean represented by the black line, the modified 2.5% confidence lower bound (lb) represented by the red line, and the 97.5% confidence upper bound (ub) represented by the blue line.
Figure B.19: Time series of $R_{eff}$: no SIA
Figure B.20: Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 70% vaccination coverage.
Figure B.21: Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 80% vaccination coverage.
Figure B.22: Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 90% vaccination coverage.
Figure B.23: Time series of $R_{eff}$: SIA target children ages 9 months old to 5 years old and achieves 95% vaccination coverage
Figure B.24: Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 70% vaccination coverage.
Figure B.25: Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 80% vaccination coverage.
Figure B.26: Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 90% vaccination coverage
Figure B.27: Time series of $R_{eff}$: SIA target children ages 9 months old to 10 years old and achieves 95% vaccination coverage
Figure B.28: Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 70% vaccination coverage.
Figure B.29: Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 80% vaccination coverage.
Figure B.30: Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 90\% vaccination coverage
Figure B.31: Time series of $R_{eff}$: SIA target children ages 9 months old to 15 years old and achieves 95% vaccination coverage
Figure B.32: Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 70% vaccination coverage
Figure B.33: Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 80% vaccination coverage
Figure B.34: Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 90% vaccination coverage
Figure B.35: Time series of $R_{eff}$: SIA target children ages 9 months old to 20 years old and achieves 95% vaccination coverage
Chapter 3 Supplementary Results

C.1 Survey Defined Age Group Bias

To tease apart the importance of finely resolved age bins in different age groups, we explored the impact of different survey defined age groupings (AG) on parameter estimation. We examined 21 age grouping schemes in four stages; grouping between ages 1 and 10, groupings between ages 10 and 20, groupings between ages 20 and 50, and finally groupings among reproductive ages 15 to 50. We used perfect seroprevalence data of each age group, i.e. the average proportion seropositive per defined age group assigned to the mid-age, and kept highly resolved point estimates for seropositivity of each age in years from the rest of the age range. After assessing biases from survey defined AG schemes based on perfect seroprevalence data by age group, we chose four final AG schemes to reflect the fewest number of age groups and the smallest bias, and then estimated parameters using perfect seroprevalence data of each age group. The following subsections explore estimation bias from all 21 AG schemes as listed here by respective age ranges: ‘Age Groupings 1-10 years old’ assesses AG1-AG4, ‘Age Groupings 10-20 years old’ assesses AG5-AG8, ‘Age Groupings 20-50 years old’ assesses AG9-AG12, ‘Age Groupings of Reproductive Ages 15-50 years old’ assesses ReproAG1-ReproAG5, and ‘Final Four Age Groupings’ assesses AG13-AG16.
C.1.1 Age Groupings 1-10 years old

First, we looked at 4 age grouping schemes that grouped ages 1 to 10 years, and kept perfect seroprevalence data at every age in years between 10 and 50 years old (represented as (10:50]). Figures C.1, C.2, and C.3 show the results of epidemiological parameter estimates using fractional polynomials for three populations with assumed $R_0$ values of 3, 5, and 7 respectively. For grouping ages 1 to 10 years the peaks in the derivative of the proportion seropositive by age curve, as seen in Figure 3.10a, play an important role in understanding the effect of age group bias on estimates of the ASFOI. In a growing population, grouping ages that are less than these peaks, say a 1-5 year age group when $R_0$ was 5, resulted in an underestimate of the derivative of proportion seropositive, $\hat{\pi}'(a)$, and an overestimate in the proportion seronegative, $(1 - \hat{\pi}(a))$ (Figure 3.10). These two biases together resulted in a large underestimate of the ASFOI in this age group. To make up for this bias, the ASFOI in the next age group, say 5-10 years old for an $R_0$ of 5, was overestimated. In other words, the magnitude of underestimation in the 1-5 year age group drove the ASFOI biases for the other ages upwards. This is why in age grouping schemes #11 and #12 (i.e. AC11 and AG12), the ASFOI was generally overestimated compared to the best analytic estimate, in turn $R_0$ was overestimated for all three populations (Figures C.1-C.3). The result of these biases on CRS was not consistent, the CRS incidence rate was underestimated when the population $R_0$ was 3, generally not affected when $R_0$ was 5, and overestimated when population $R_0$ was 7.

Figures C.4, C.5, and C.6 show the results of epidemiological parameter estimates using penalized regression splines for age groupings between 1 to 10 years for three populations with assumed $R_0$ values of 3, 5, and 7 respectively. Just like fractional polynomials, in a growing population, grouping ages that are younger than the peak age of $\hat{\pi}'(a)$ resulted in a large underestimate of the ASFOI in this age group (Figure 3.10). As a result, $R_0$ was also underestimated (see AG11 in Figures C.4-C.6). Grouping ages that are older than the peak age of $\hat{\pi}'(a)$ means $\hat{\pi}'(a)$ was overestimated, which drove ASFOI and $R_0$ to also be overestimated (see AG10 in Figures C.4-C.6). When we grouped ages 1-5 and 5-10 (see AG12 in Figures C.4-C.6), these two biases counteracted one another and the estimate of $R_0$ was similar to the best analytic estimate. The result of age groupings [1,5,10] on estimates of the CRS incident rate depended on population transmission;
Figure C.1: \( R_0 = 3 \), Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.

CRS incident rate was overestimated when \( R_0 \) was 3 or 5, and underestimated with assumed \( R_0 \) was 7. Despite these biases, the overall effect on estimates of the AFOI, \( R_0 \), and CRS incident was minimal in magnitude.

Figures C.7, C.8, and C.9 show the results of epidemiological parameter estimates using the two-age group piecewise constant analytic method for age groupings between ages 1 to 10 years for three populations with assumed \( R_0 \) values of 3, 5, and 7 respectively. Grouping ages between 1-5 or 5-10 (AC9, AC10, AC11) caused the younger age FOI, \( \tilde{\lambda}_y \), and \( R_0 \) to be overestimated compared to the best analytic estimate (Figures C.7, C.9). Grouping both ages 1-5 and 5-10 (AG12) therefore resulted in even larger overestimates of younger age FOI, \( \tilde{\lambda}_y \), and \( R_0 \); however \( R_0 \) estimates were moving towards the true parameter values (Figures C.7, C.9). The two-age group piecewise constant
Figure C.2: $R_0 = 5$, Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.

analytic method is compensatory, such that overestimates in younger age FOI, $\bar{\lambda}_y$ (compared to best analytic estimate of $\bar{\lambda}_y$), caused underestimates in older age FOI, $\bar{\lambda}_o$ (compared to best analytic estimate of $\bar{\lambda}_o$), of similar magnitude. Figures C.7-C.9 therefore showed drops in the CRS incidence rate across age grouping schemes, however estimates were biases towards the true parameter values for each population.
Figure C.3: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.4: $R_0 = 3$, Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2 column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.5: $R_0 = 5$, Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,0000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.6: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.7: \( R_0 = 3 \), Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using two-age group piecewise constants analytic method. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.8: \( R_0 = 5 \), Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using two-age group piecewise constants analytic method. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.9: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 1-10 years using two-age group piecewise constants analytic method. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
C.1.2 Age Groupings 10-20 years old

Second, we explored 4 age grouping schemes that group ages 10 to 20 years, and kept perfect seroprevalence data at every age in years between 1 and 10 years old (represented as [1:10]), and also between 20 to 50 years old (represented as (20:50]). Across these age groupings, we found agreement in direction of bias for all analytic methods and assumed population $R_0$ values (Figures C.10 - C.18). For 10 to 20 year age groupings, the overestimate in the proportion seronegative, $(1 - \hat{\pi}(a))$, (Figure 3.10b) in a growing population mostly drove the bias in the ASFOI estimates downwards using fractional polynomial and penalized regression splines analytic methods. The same was true using the piecewise constant analytic method; underestimation of proportion sero-positive caused the younger age FOI, $\bar{\lambda}_y$, to be underestimated. As a result, as age groups widened between the ages of 10 and 20, estimates of $R_0$ decreased moving further away from the best analytic estimate and true parameter value (Figures C.10 - C.18). Additionally, as age groups widened between 10 and 20 years old, estimates of the CRS incidence rates increased away from the best analytic estimate due to a larger proportion of susceptible individuals at early reproductive ages (Figures C.10 - C.15). In the piecewise constant method, the increase in CRS was also due to overestimates in the older age force of infection, $\bar{\lambda}_o$, to compensate for underestimated in younger age force of infection, $\bar{\lambda}_y$ (Figures C.16 - C.15). We conclude that it was important to break up the 10 to 20 years olds to reduce age group bias, and the more age groups included between 10 and 20 the lower the magnitude of mis-estimation of epidemiological parameters of $R_0$ and the CRS incidence rate.

The literature shows that 13 years old may be theoretically a good choice because this is when children often leave education and age contact patterns change (Vynnycky et al., 2016), however, for rubella, the 15 years old is also an important break point to classify reproductive ages from non-reproductive ages.
Figure C.10: $R_0 = 3$, Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using fractional polynomials analytic method. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the grey vertical lines represent the age groups, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.11: \( R_0 = 5 \), Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.12: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.13: $R_0 = 3$, Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the grey vertical lines represent the age groups, the black vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.14: $R_0 = 5$, Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the grey vertical lines represent the age groups, the black dots represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.15: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.16: $R_0 = 3$, Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the grey vertical lines represent the age groups, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.17: \( R_0 = 5 \), Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.18: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 10-20 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
C.1.3 Age Groupings 20-50 years old

Third, we looked at 4 age grouping schemes that group ages 20 to 50 years, and kept perfect seroprevalence data at every age in years between 1 and 20 years old (represented as [1:20)). Figures C.19, C.20, and C.21 show the results of epidemiological parameter estimates using fractional polynomials for three populations with assumed \( R_0 \) values of 3, 5, and 7 respectively. For fractional polynomials, the overestimate in \( \hat{\pi}'(a) \) at late ages drove the bias in ASFOI upwards resulting in overestimates of \( R_0 \) compared to the analytic best estimate, but closer to the true value. An increase in ASFOI meant lower proportion seronegative, or fewer individuals susceptible to infection in reproductive age, and resulted in a lower CRS incidence rate. The exception was the population with an \( R_0 \) of 3, which showed a drop in the estimated CRS incidence rate with 5 years age groups between 20 and 50 years old, but an increase in the CRS incidence rate as age groups widen (Figure C.19). This occurred because there were more estimated susceptible individuals at the beginning of the reproductive ages. In general, using fractional polynomials, grouping ages 20 to 50 years old into one large age group resulted in mis-estimation of epidemiological parameters toward the true parameter values.

Figures C.22, C.23, and C.24 show the results of epidemiological parameter estimates using penalized regression splines for age groupings between 20 to 50 years for three populations with assumed \( R_0 \) values of 3, 5, and 7 respectively. For age groupings 25 to 50 and 20 to 50 (AG3 and AG4), the overestimate in \( (1-\hat{\pi}(a)) \) at late ages drove the bias in the ASFOI downwards, resulting in underestimates of \( R_0 \) away from truth. However, the change in \( R_0 \) estimate was very small ranging from 0.03 to 0.06 across different \( R_0 \) values. Across the three different populations, grouping ages 20 to 50 years into one large age class underestimated the CRS incidence up to 7 per 100,000 live births (Figure C.22). Therefore, age groupings that include at least one five year age class between 20 and 25 was important to prevent underestimates of CRS incidence rates when using the penalized regression splines.

Figures C.25, C.26, and C.27 show the results of epidemiological parameter estimates using piecewise constant analytic method for age groupings between 20 to 50 years for three populations with
Figure C.19: $R_0 = 3$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.

Assumed $R_0$ values of 3, 5, and 7 respectively. Age grouping bias for piecewise constant was similar to the biases seen using fractional polynomials; the ASFOI was overestimated at later ages as age groups widen. The result was overestimated $R_0$ values compared to the analytic best estimate, but closer to the true value. The effect on the estimate of the CRS incidence rate depends on the population rate of transmission; the older age FOI, $\bar{\lambda}_o$, was overestimated for all populations but the effect on CRS depends on how fast susceptible individuals are depleted. Conclusions are similar as fractional polynomials method, that the piecewise constant approach results in mis-estimation of $R_0$ towards to true parameter values from grouping ages 20 to 50 years. However, the effect of grouping ages 20 to 50 into one age group on CRS depends on the assumed population $R_0$; estimates were generally biased away from both true value and best analytic estimate when $R_0$ was 3 and 5, and underestimated toward true parameter value when $R_0$ was 5.
Figure C.20: $R_0 = 5$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.21: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.22: \( R_0 = 3 \), Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.23: $R_0 = 5$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the grey vertical lines represent the age groups, the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.24: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the grey vertical lines represent the reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.25: $R_0 = 3$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG; the grey vertical lines represent the age groups, the black dots represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.26: $R_0 = 5$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.27: $R_0 = 7$, Epidemiological parameter estimates for Age Groupings (AG) of ages 20-50 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
C.1.4 Age Groupings of Reproductive Ages 15-50 years old

Given that many rubella serological surveys sample reproductive ages only, we explored 5 age grouping schemes that grouped reproductive ages 15 to 50 years, and did not ‘sample’ from any other ages. Figures C.28, C.29, and C.30 show the results of epidemiological parameter estimates using fractional polynomials for three populations with assumed $R_0$ values of 3, 5, and 7 respectively. Figures C.31, C.32, and C.33 show the results of epidemiological parameter estimates using penalized regression splines for age groupings that group reproductive ages 15 to 50 years for three populations with assumed $R_0$ values of 3, 5, and 7 respectively. Figures C.34, C.35, and C.36 show the results of epidemiological parameter estimates using piecewise constant analytic method for age groupings that group reproductive ages 15 to 50 years for three populations with assumed $R_0$ values of 3, 5, and 7 respectively. Overall, we generally found that grouping reproductive ages into as few as three ages groups make little difference in CRS estimates across analytic methods (ReproAC1-ReproAC3 vs best analytic estimate among repro ages), with the exception of Figure C.29 and C.30. However, seroprevalence data with only 2 age groups generally biased CRS incidence rates downwards using fractional polynomials and piecewise constant analytic methods (with the exception of Figure C.29). And, penalized regression splines analytic method cannot estimate the ASFOI or the CRS incidence rate from fewer than three age groups given limits in degrees of freedom.
Figure C.28: $R_0 = 3$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.8a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 column 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.29: $R_0 = 5$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.7a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.30: $R_0 = 7$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.9a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.31: \( R_0 = 3 \), Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.8b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.32: $R_0 = 5$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.7b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.33: $R_0 = 7$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.9b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.34: $R_0 = 3$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15–50 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.8c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.35: $R_0 = 5$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.7c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
Figure C.36: $R_0 = 7$, Epidemiological parameter estimates for Reproductive Age Groupings (ReproAG) of ages 15-50 years using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every reproductive age in years 15 to 50. These are referred to as ‘best analytic estimates among repro ages’ and this figure is a repetition of Figure 3.9c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 columns 1-2 show the proportion seropositive and ASFOI for 5 Reproductive Age Groupings (ReproAG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 3 shows the true and estimated CRS incidence rate by reproductive age for each ReproAG. The figure in row 2 columns 4 shows the dot plot of estimated population CRS incidence per 100,000 live births for each ReproAG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate among repro ages.
C.1.5 Final Four Age Groupings

Based on our findings of age grouping bias, as detailed above, we selected four final age grouping (AG) schemes to both minimize the number of age groups and age group bias:

- \( AG_{13} : [1, 5, 10, 13, 15, 20, 25, 50] \),
- \( AG_{14} : [1, 5, 10, 13, 15, 20, 50] \),
- \( AG_{15} : [1, 5, 10, 15, 20, 25, 50] \),
- \( AG_{16} : [1, 5, 10, 15, 20, 50] \).

It is first important to note that all final age grouping schemes biased epidemiological parameter estimates away from the best analytic estimates described above; the overall direction and magnitude of bias depended on the analytic method and true age immunity profile of the population.

Figures C.37, C.38, and C.39 show the results of epidemiological parameter estimates using fractional polynomials analytic method for the final four age grouping schemes for three populations with assumed \( R_0 \) values of 3, 5, and 7 respectively. Under the fractional polynomials analytic method, including the 10-13 age group was preferred as it generally resulted in a higher ASFOI curve compared to best estimate from fractional polynomial method, and biases in estimation of \( R_0 \) and CRS incidence rate were toward true parameter values. Also, not including the 20-25 year age group tended to bias the ASFOI, \( R_0 \), and the CRS incidence rate toward true parameter values, although minimally.

Figures C.40, C.41, and C.42 show the results of epidemiological parameter estimates using the two-age group piecewise constant analytic method for the final four age grouping schemes for three populations with assumed \( R_0 \) values of 3, 5, and 7 respectively. Across these final four age grouping schemes, we found the more age classes the better parameter estimates when analyzing seroprevalence data. Age grouping schemes further underestimated \( R_0 \) away from the true parameter value and best estimate, and further overestimated CRS incidence rates away from the true parameter value and best estimate as age groups decreased.
Figure C.37: $R_0 = 3$, Epidemiological parameter estimates for the final four Age Groupings (AG) using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.

Figures C.43, C.44, and C.45 show the results of epidemiological parameter estimates using penalized regression splines analytic method for the final four age grouping schemes for three populations with assumed $R_0$ values of 3, 5, and 7 respectively. As shown the analysis on bias in age grouping 20-50 above, it was very important to include the 20-25 age group in order to estimate CRS incidence, however the effect of not including the 20-25 age class was minimal on $R_0$ estimates. When estimating $R_0$, it was important to include the 10-13 age class; however, including the 20-25 age group becomes increasingly less necessary as the true value of $R_0$ increases. Regardless of these biases, penalized regression splines still estimated the most accurate parameters compared to two-age group piecewise constants and fractional polynomials.
Figure C.38: \( R_0 = 5 \), Epidemiological parameter estimates for the final four Age Groupings (AG) using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.39: $R_0 = 7$, Epidemiological parameter estimates for the final four Age Groupings (AG) using fractional polynomials. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5a, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.40: $R_0 = 3$, Epidemiological parameter estimates for the final four Age Groupings (AG) using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.41: $R_0 = 5$, Epidemiological parameter estimates for the final four Age Groupings (AG) using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.42: $R_0 = 7$, Epidemiological parameter estimates for the final four Age Groupings (AG) using two-age group piecewise constants. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5c, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.43: $R_0 = 3$, Epidemiological parameter estimates for the final four Age Groupings (AG) using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.4b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.44: \( R_0 = 5 \), Epidemiological parameter estimates for the final four Age Groupings (AG) using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.3b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and \( R_0 \) for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.
Figure C.45: $R_0 = 7$, Epidemiological parameter estimates for the final four Age Groupings (AG) using penalized regression splines. The figure in row 1, column 1 shows the true and estimated proportion seropositive and ASFOI assuming perfect data at every age in years. These are referred to as ‘best analytic estimates’ and this figure is a repetition of Figure 3.5b, shown again here for ease of distinguishing analytic method bias and age grouping bias. The figures in row 1, columns 2-4, and row 2 column 1 show the proportion seropositive and ASFOI for 4 Age Groupings (AG); the grey vertical lines represent the age groups, the black dots represent perfect seroprevalence data or the mean proportion seropositive for each age group, the red lines represent the true proportion seropositive and ASFOI, and the dashed blue lines represent the estimated proportion seropositive and ASFOI. The figure in row 2, column 2 shows the true and estimated CRS incidence rate by reproductive age for each AG. The figures in row 2 columns 3 and 4 show dot plots of estimated population CRS incidence per 100,000 live births and $R_0$ for each AG; the red vertical lines represent the true parameter value and the vertical orange dashed lines represent the best analytic estimate.