An Integrative Metaregression Framework for Descriptive Epidemiology
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>3</td>
</tr>
<tr>
<td>List of Figures</td>
<td>7</td>
</tr>
<tr>
<td>List of Tables</td>
<td>11</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>13</td>
</tr>
<tr>
<td>Abraham D. Flaxman, Theo Vos, and Christopher J. L. Murray</td>
<td></td>
</tr>
<tr>
<td>1.1 A motivating example: Parkinson’s disease</td>
<td>17</td>
</tr>
<tr>
<td>1.2 From systematic review to metaregression</td>
<td>21</td>
</tr>
<tr>
<td>1.3 History of generic disease modeling</td>
<td>24</td>
</tr>
<tr>
<td>2 Theory</td>
<td>29</td>
</tr>
<tr>
<td>Abraham D. Flaxman</td>
<td></td>
</tr>
<tr>
<td>2.1 A motivating example: Schizophrenia prevalence</td>
<td>32</td>
</tr>
<tr>
<td>2.2 Binomial model</td>
<td>33</td>
</tr>
<tr>
<td>2.3 Beta-binomial model</td>
<td>36</td>
</tr>
<tr>
<td>2.4 Poisson model</td>
<td>38</td>
</tr>
<tr>
<td>2.5 Negative-binomial model</td>
<td>40</td>
</tr>
<tr>
<td>2.6 Transformed normal models</td>
<td>41</td>
</tr>
<tr>
<td>2.7 Lower-bound data model</td>
<td>44</td>
</tr>
<tr>
<td>2.8 Quantification of uncertainty</td>
<td>46</td>
</tr>
<tr>
<td>2.9 Comparison</td>
<td>47</td>
</tr>
<tr>
<td>2.10 Summary and future work</td>
<td>50</td>
</tr>
</tbody>
</table>
CONTENTS

3 Age pattern models
Abraham D. Flaxman
3.1 Definition of spline models ........................................ 52
3.2 Choosing knots ......................................................... 54
3.3 Penalized spline models ............................................. 55
3.4 Augmenting the spline model ....................................... 57
3.5 Summary and future work .......................................... 58

4 Expert priors on age patterns
Abraham D. Flaxman
4.1 Priors on level ......................................................... 59
4.2 Priors on monotonicity .............................................. 62
4.3 Priors are not just for splines ..................................... 63
4.4 Empirical priors on age patterns ................................ 64
4.5 Summary and future work .......................................... 65

5 Statistical models for heterogeneous age groups
Abraham D. Flaxman
5.1 Overlapping age-group data ......................................... 69
5.2 Midpoint model ......................................................... 70
5.3 Disaggregation model ................................................ 72
5.4 Midpoint model with group width covariate .................. 73
5.5 Age-standardizing and age-integrating models ................ 75
5.6 Model comparison ..................................................... 77
5.7 Summary and future work .......................................... 79

6 Covariate modeling
Abraham D. Flaxman
6.1 Cross-walk fixed effects to explain bias ....................... 82
6.2 Predictive fixed effects to improve out-of-sample estimation 85
6.3 Fixed effects to explain variance ................................ 87
6.4 Random effects for spatial variation ............................ 87
6.5 Covariates and consistency ........................................ 89
6.6 Summary and future work .......................................... 90

7 Prevalence estimates from other data types
Abraham D. Flaxman
7.1 A motivating example: population dynamics .................. 94
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Cross-walking with fixed effects: anxiety disorders</td>
<td>Amanda Baxter, Jed Blore, Abraham D. Flaxman, Theo Vos, and Harvey Whiteford</td>
</tr>
<tr>
<td>15</td>
<td>Improving out-of-sample prediction: liver cirrhosis</td>
<td>Ali Mokdad, Abraham D. Flaxman, Hannah M. Peterson, Christopher J. L. Murray, and Mohsen Naghavi</td>
</tr>
<tr>
<td>16</td>
<td>Risk factors: fruit consumption</td>
<td>Stephen S. Lim, Hannah M. Peterson, and Abraham D. Flaxman</td>
</tr>
<tr>
<td>17</td>
<td>The compartmental model: end-stage renal disease</td>
<td>Sarah K. Wolf, Abraham D. Flaxman, Mohsen Naghavi, and Giuseppe Remuzzi</td>
</tr>
<tr>
<td>18</td>
<td>Knot selection in compartmental spline models: osteoarthritis of the knee</td>
<td>Marita Cross, Damian Hoy, Theo Vos, Abraham D. Flaxman, and Lyn March</td>
</tr>
<tr>
<td>19</td>
<td>Expert priors in compartmental models: bipolar disorder</td>
<td>Alizé Ferrari, Abraham D. Flaxman, Hannah M. Peterson, Theo Vos, and Harvey Whiteford</td>
</tr>
<tr>
<td>20</td>
<td>Cause-specific mortality rates: alcohol dependence</td>
<td>Theo Vos, Jed Blore, Abraham D. Flaxman, Hannah M. Peterson, and Juergen Rehm</td>
</tr>
<tr>
<td>21</td>
<td>Conclusion</td>
<td>Abraham D. Flaxman, Christopher J. L. Murray, and Theo Vos</td>
</tr>
<tr>
<td></td>
<td>A GBD Study 2010 spatial hierarchy</td>
<td></td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Systematic review data for Parkinson’s disease.</td>
<td>18</td>
</tr>
<tr>
<td>1.2</td>
<td>Estimates of age-specific rates for Parkinson’s disease.</td>
<td>20</td>
</tr>
<tr>
<td>2.1</td>
<td>Forest plot of 7 alternative models for meta-analysis of schizophrenia.</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Funnel plot for the binomial rate model.</td>
<td>35</td>
</tr>
<tr>
<td>2.3</td>
<td>Posterior predictive check for binomial model fitted to adult male schizophrenia data.</td>
<td>37</td>
</tr>
<tr>
<td>2.4</td>
<td>Probability density for the beta distribution.</td>
<td>38</td>
</tr>
<tr>
<td>2.5</td>
<td>Funnel plots and posterior predictive check for the beta-binomial model.</td>
<td>39</td>
</tr>
<tr>
<td>2.6</td>
<td>Funnel plots and posterior predictive check for the negative-binomial model.</td>
<td>42</td>
</tr>
<tr>
<td>2.7</td>
<td>Funnel plots and posterior predictive check for the offset log-normal model.</td>
<td>45</td>
</tr>
<tr>
<td>3.1</td>
<td>All-cause mortality as a function of age.</td>
<td>52</td>
</tr>
<tr>
<td>3.2</td>
<td>Spline interpolation of simulated data.</td>
<td>54</td>
</tr>
<tr>
<td>3.3</td>
<td>Penalized splines with a range of smoothing parameters.</td>
<td>56</td>
</tr>
<tr>
<td>4.1</td>
<td>An informative prior on the level of $h(a)$.</td>
<td>60</td>
</tr>
<tr>
<td>4.2</td>
<td>An informative prior on the upper and lower bounds of the age-specific hazard function $h(a)$.</td>
<td>61</td>
</tr>
<tr>
<td>4.3</td>
<td>An informative prior that the age pattern is increasing or decreasing across an age range.</td>
<td>63</td>
</tr>
<tr>
<td>5.1</td>
<td>Mean and spread of age groups collected from a systematic review of atrial fibrillation.</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>Systematic review data for atrial fibrillation.</td>
<td>70</td>
</tr>
</tbody>
</table>
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>The midpoint model applied to homogeneous and heterogeneous age groups</td>
<td>72</td>
</tr>
<tr>
<td>5.4</td>
<td>The disaggregation approach applied to homogeneous and heterogeneous age groups</td>
<td>74</td>
</tr>
<tr>
<td>5.5</td>
<td>The midpoint-covariate model applied simulated data</td>
<td>75</td>
</tr>
<tr>
<td>5.6</td>
<td>The age-standardizing model applied to simulated data</td>
<td>76</td>
</tr>
<tr>
<td>5.7</td>
<td>A comparison of 4 models for heterogeneous age groups</td>
<td>77</td>
</tr>
<tr>
<td>6.1</td>
<td>Simulated data set where different measurement techniques yield systematically different values</td>
<td>83</td>
</tr>
<tr>
<td>6.2</td>
<td>Data on bipolar disorder collected in systematic review</td>
<td>84</td>
</tr>
<tr>
<td>6.3</td>
<td>Comparison of reference value choice on incidence estimates</td>
<td>86</td>
</tr>
<tr>
<td>7.1</td>
<td>A single-compartment model with inflow $b$ and outflow $m$</td>
<td>94</td>
</tr>
<tr>
<td>7.2</td>
<td>Stock and flow in a one-compartment model as a function of time</td>
<td>96</td>
</tr>
<tr>
<td>7.3</td>
<td>The two-compartment model of process for disease in a population</td>
<td>97</td>
</tr>
<tr>
<td>7.4</td>
<td>Consistent disease parameters for a condition</td>
<td>102</td>
</tr>
<tr>
<td>7.5</td>
<td>Consistent disease parameters for a condition with excess mortality $h_f$ increasing linearly as a function of age</td>
<td>103</td>
</tr>
<tr>
<td>7.6</td>
<td>Examples of different age-specific rates and the resulting prevalence curves</td>
<td>104</td>
</tr>
<tr>
<td>8.1</td>
<td>The results of drawing 100 samples from a ball in two dimensions with MCMC using the Gibbs step method</td>
<td>110</td>
</tr>
<tr>
<td>8.2</td>
<td>The results of drawing 100 samples from an ellipse in two dimensions with MCMC using the MH step method</td>
<td>113</td>
</tr>
<tr>
<td>8.3</td>
<td>The results of drawing 100 samples from an ellipse in two dimensions with MCMC using the AM step method, after 20,000 iterations of “burn-in.”</td>
<td>115</td>
</tr>
<tr>
<td>9.1</td>
<td>Systematic review data for cocaine dependence</td>
<td>123</td>
</tr>
<tr>
<td>9.2</td>
<td>Prevalence estimates of cocaine dependence using spline models with varying numbers of knots</td>
<td>124</td>
</tr>
<tr>
<td>9.3</td>
<td>Prevalence estimates of cocaine dependence using a smoothing parameter</td>
<td>125</td>
</tr>
<tr>
<td>10.1</td>
<td>Systematic review data for premenstrual syndrome</td>
<td>128</td>
</tr>
<tr>
<td>10.2</td>
<td>Four estimates of age-specific PMS prevalence with different priors on prevalence in young and old ages</td>
<td>129</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>10.3</td>
<td>Estimate of age-specific PMS prevalence for spline models with a variety of knots.</td>
<td>131</td>
</tr>
<tr>
<td>10.4</td>
<td>Estimates of age-specific PMS prevalence for spline models with a variety of monotonicity priors.</td>
<td>132</td>
</tr>
<tr>
<td>10.5</td>
<td>Prevalence estimates for women in Western Europe with PMS.</td>
<td>133</td>
</tr>
<tr>
<td>11.1</td>
<td>Systematic review data for pancreatitis with estimates.</td>
<td>136</td>
</tr>
<tr>
<td>11.2</td>
<td>Comparison of pancreatitis incidence estimates.</td>
<td>137</td>
</tr>
<tr>
<td>11.3</td>
<td>Systematic review data of atrial fibrillation.</td>
<td>140</td>
</tr>
<tr>
<td>11.4</td>
<td>Comparison of prevalence estimates for atrial fibrillation using an age-standardizing model and midpoint model.</td>
<td>141</td>
</tr>
<tr>
<td>11.5</td>
<td>Comparison of incidence estimates for atrial fibrillation using an age-standardizing model and midpoint model.</td>
<td>142</td>
</tr>
<tr>
<td>11.6</td>
<td>Estimates of atrial fibrillation using an age-standardizing compartmental model.</td>
<td>143</td>
</tr>
<tr>
<td>11.7</td>
<td>Comparison of prevalence and incidence estimates for atrial fibrillation using an age-standardizing model and midpoint model.</td>
<td>143</td>
</tr>
<tr>
<td>13.1</td>
<td>Prevalence data from systematic review of anti-HCV in (a) Egypt and (b) Jordan.</td>
<td>147</td>
</tr>
<tr>
<td>13.2</td>
<td>Estimates of anti-HCV prevalence in North Africa and Middle East region and in the countries Egypt and Jordan.</td>
<td>148</td>
</tr>
<tr>
<td>13.3</td>
<td>The intercept shift of anti-HCV prevalence with different priors on global heterogeneity.</td>
<td>149</td>
</tr>
<tr>
<td>14.1</td>
<td>Systematic review data of anxiety disorders.</td>
<td>152</td>
</tr>
<tr>
<td>14.2</td>
<td>Comparison of prevalence estimates of anxiety disorders using different methods.</td>
<td>153</td>
</tr>
<tr>
<td>15.1</td>
<td>Systematic review data for cirrhosis.</td>
<td>156</td>
</tr>
<tr>
<td>15.2</td>
<td>Relationship between cirrhosis prevalence data collected in systematic review and age-standardized death rate of cirrhosis.</td>
<td>157</td>
</tr>
<tr>
<td>15.3</td>
<td>Cirrhosis prevalence estimates and available data for the USA and Egypt.</td>
<td>158</td>
</tr>
<tr>
<td>16.1</td>
<td>Fruit consumption estimates in USA.</td>
<td>160</td>
</tr>
<tr>
<td>16.2</td>
<td>Fruit consumption estimates in Greece and Iceland.</td>
<td>161</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>16.3</td>
<td>Comparison of posterior distribution estimates of fruit consumption.</td>
<td>162</td>
</tr>
<tr>
<td>17.1</td>
<td>Systematic review data for end-stage renal disease treated with</td>
<td>166</td>
</tr>
<tr>
<td></td>
<td>dialysis.</td>
<td></td>
</tr>
<tr>
<td>17.2</td>
<td>Comparison of epidemiological parameter estimates for end-stage</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>renal disease treated with dialysis using the compartmental and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spline models.</td>
<td></td>
</tr>
<tr>
<td>17.3</td>
<td>Comparison of the regional age-standardized prevalence estimates</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td>using compartmental and spline models for end-stage renal disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td>treated with dialysis.</td>
<td></td>
</tr>
<tr>
<td>17.4</td>
<td>Comparison of with-condition mortality estimates for end-stage</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>renal disease treated with dialysis using a compartmental model,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a spline model, and a spline model with a smoothing parameter.</td>
<td></td>
</tr>
<tr>
<td>18.1</td>
<td>Comparison of osteoarthritis of the knee estimates using a</td>
<td>172</td>
</tr>
<tr>
<td></td>
<td>compartmental model with different knot selections.</td>
<td></td>
</tr>
<tr>
<td>18.2</td>
<td>Comparison of osteoarthritis of the knee estimates with and</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>without a prior restricting the age of onset.</td>
<td></td>
</tr>
<tr>
<td>19.1</td>
<td>Systematic review data for bipolar disorder.</td>
<td>177</td>
</tr>
<tr>
<td>19.2</td>
<td>Comparison of prevalence estimates for bipolar disorder using</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>differing priors for the age of onset in a compartmental model.</td>
<td></td>
</tr>
<tr>
<td>19.3</td>
<td>Comparison of estimates for bipolar disorder using a compartmental</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>model with different priors on the upper age limit of incidence.</td>
<td></td>
</tr>
<tr>
<td>19.4</td>
<td>Comparison of estimates for bipolar disorder using a compartmental</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>model with different priors on remission.</td>
<td></td>
</tr>
<tr>
<td>20.1</td>
<td>Systematic review data for alcohol dependence.</td>
<td>182</td>
</tr>
<tr>
<td>20.2</td>
<td>The 2-compartment mechanistic model with the excess mortality</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td>split into 2 hazards.</td>
<td></td>
</tr>
<tr>
<td>20.3</td>
<td>Comparison of estimates for alcohol dependence when using cause-</td>
<td>184</td>
</tr>
<tr>
<td></td>
<td>specific mortality data as a lower bound ($h_{f'w}$ unrestricted) or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a direct estimate of the product of prevalence and excess mortality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>($h_{f'w} = 0$)</td>
<td></td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Mean results of holdout cross-validation of rate models for 100 replicates for out-of-sample prediction, ordered by increasing median absolute error (MAE).</td>
<td>49</td>
</tr>
<tr>
<td>5.1</td>
<td>Median results for 100 replicates of the simulation study comparing age-specific rate estimates from 5 models of age-grouped data.</td>
<td>79</td>
</tr>
<tr>
<td>13.1</td>
<td>Estimates of the country-level random effects for anti-HCV prevalence.</td>
<td>147</td>
</tr>
<tr>
<td>13.2</td>
<td>Anti-HCV age-standardized prevalence estimates from a hierarchical random-effects spline model with differing priors on global heterogeneity.</td>
<td>150</td>
</tr>
<tr>
<td>16.1</td>
<td>Random-effect estimates from hierarchical spline models of fruit consumption with differing rate models.</td>
<td>163</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Abraham D. Flaxman, Theo Vos, and Christopher J. L. Murray
This book, *An Integrative Metaregression Framework for Descriptive Epidemiology*, is a full-length treatment of new meta-analytic methods for descriptive epidemiology. From first principles, it develops the integrative systems model that constitutes the theoretical foundation of morbidity estimation in burden of disease studies like the Global Burden of Disease, Injuries, and Risk Factors (GBD 2010) Study 2010. The estimation approach relies on producing age-specific prevalence estimates of the nonfatal outcomes of a vast array of diseases, injuries, and risk factors. As part of the GBD Study 2010, we have developed a Bayesian metaregression tool specifically for this purpose. This tool estimates a generalized negative-binomial model for all the epidemiological data with various types of fixed and random effects. These include age fixed effects, fixed effects for covariates that predict country variation in the quantity of interest, fixed effects that predict variation across studies due to attributes of the study protocols, and superregion, region, and country random intercepts. The tool uses Bayesian inference of the parameters to sample from the joint posterior distribution of the model, incorporating all relevant descriptive epidemiological data. This approach is new, but the line of research builds on work in generic disease modeling that has been in use for almost 20 years in global health epidemiology. However, until now, the descriptions of the models and the methods have been scattered throughout the scientific literature in a loose collection of journal articles, burden of disease reports, and operations manuals.

This book substantially extends the previous modeling efforts for YLD estimation in burden of disease estimation by formally connecting a system dynamics model of disease progression to a statistical model of the epidemiological rates calculated in descriptive epidemiological research and collected in a systematic review. This combination of systems dynamics modeling and statistical modeling, which we call *integrative systems modeling*, allows the model to integrate all available relevant data. Because of the advanced numerical algorithms needed to fit these complex models, chapter 8 provides the necessary background on Markov chain Monte Carlo (MCMC) and other relevant computational methods.

Experience with the results of systematic review indicates that when all available relevant data are collected, they are often very *sparse* and very *noisy*. In GBD estimation, data sparsity often means that there are whole regions of the globe for which no data are available. The sparsity of data means that predictions of prevalence need to take advantage of relationships
to covariates in the metaregression or default to the average of a region, superregion, or the world. Dealing with noisy data is an additional challenge. In the regions or countries with multiple measurements, the results are often highly heterogeneous. The degree of heterogeneity is far beyond what is expected on the basis of sampling error and indicates considerable non-sampling variance. The sources of non-sampling variance include challenges in sample design; lack of a representative sample; and implementation issues in data collection, case definitions, and diagnostic technologies. To make matters more complicated, there is true geographic variation as well.

We will also address a number of other common challenges in estimating the prevalence of nonfatal outcomes of disease:

- Based on biology, exposure, or clinical series, we may have strong priors on the age pattern of incidence or prevalence of a condition; for example, due to cumulative exposure to carcinogens, we expect the incidence of many cancers to increase with age, at least until some adult age. Another example is that the prevalence of bipolar disorder is zero in younger children.

- Published studies often use diverse age groups like 18–35 or 15 and older. For the GBD Study 2010, we needed to use data from different nonstandard age groups to generate coherent estimates for the 20 age groups in the study. Given that prevalence for most sequelae is strongly related to age, this issue is particularly important.

- For many conditions, the available studies use different case definitions. The review of diabetes prevalence studies identified 18 different case definitions in use. If all non-reference definition data are excluded, predictions can be based on only a limited number of studies. An alternative is to empirically adjust between different definitions using the overlap in available studies.

- Within regions or countries, the true prevalence for a sequela can vary enormously. The high level of hepatitis C in Egypt is an example in the Middle East and North Africa region. Such within-region heterogeneity in the true rates must be accommodated in a metaregression framework.

- Data are collected for many different outcomes, such as incidence, prevalence, remission, excess mortality, or cause-specific mortality.
The mix of data varies across diseases and across regions for a disease. All these sources provide some relevant information for estimating prevalence.

The statistical model developed in this book focuses particularly on techniques for handling sparse, noisy data while also addressing these additional challenges. The book explores statistical models for overdispersed count data, covariate modeling to explain systematic variation in epidemiological data and to increase predictive accuracy for estimates where no data are available, and age pattern modeling to systematically incorporate expert knowledge about how epidemiological rates vary as a function of age. It also develops a theory of age-group modeling to address the heterogeneity in age groups that is commonly found during systematic review.

In the first half of this book, we present the theoretical foundations of integrative systems modeling of disease in populations. The second half of the book contains a series of applications of the model to the meta-analysis of a dozen different diseases. These practical applications demonstrate how the model performs in a variety of scenarios and also how the model performs when the model assumptions are violated.
1.1  A motivating example: Parkinson’s disease

As a motivating example, we now turn to the descriptive epidemiological metaregression of Parkinson’s disease (PD). A systematic review of PD was conducted as part of the GBD Study 2010. The results of this review needed to be combined to produce estimates of disease prevalence by region, age, sex, and year. These prevalence estimates were combined with disability weights to measure years lived with disability (YLDs), which were then combined with estimates of years of life lost (YLLs) to produce estimates of the burden of PD quantified in disability-adjusted life-years (DALYs).

PD is a neurodegenerative disorder that includes symptoms of motor dysfunction, such as tremors, rigidity, and akinesia, in the early stages of the disease. As the disease develops, most patients also develop nonmotor symptoms, such as cognitive decline, dementia, autonomic failure, and disordered sleep-wake regulation. The standard definition for PD diagnosis includes at least two of four cardinal signs—resting tremor, bradykinesia, rigidity, and postural abnormalities. There is no cure or treatments to slow the progression of the disease; however, motor symptoms and disability may be improved with symptomatic therapy.

Systematic review for PD yielded 782 data points that met the inclusion criteria: 660 prevalence, 99 incidence, and 13 standardized mortality ratio data points. Prevalence is measured as the ratio of the number of individuals with the condition to the number of individuals in the total population; incidence is the rate at which an individual acquires the condition; and the standardized mortality ratio is the ratio of the mortality rate of individuals with the condition to the mortality rate of the total population. A separate analysis of mortality due to PD was carried out as part of GBD 2010 and provided 1638 additional data points of cause-specific mortality by region, age, and sex for the years 1990, 2005, and 2010. Prevalence is measured as a percent, standardized mortality ratio is a unitless number, and incidence and cause-specific mortality is measured in 10,000 person years (PY). It is important to note that in this example, cause-specific mortality means that the person issuing a death certificate deemed the person to have died from the disease as the underlying cause and not simply with it. This subtle distinction is elaborated on in section 2.7 and demonstrated further in chapter 20.
Even when restricting the data to a specific geographic region, such as Western Europe, the data remain noisy and heterogeneous, as seen in figure 1.1. This figure shows a horizontal bar for each data point, where the left and right endpoints depict the start and end ages of the age interval measured, and the position of the bar on the $y$-axis indicates the value of the measurement. Section 5.1 describes our approach to making robust estimates in the face of such heterogeneous levels and overlapping age groups.

Figure 1.1. Data points from systematic review of descriptive epidemiology of Parkinson’s disease, showing data from Western Europe on (a) prevalence, (b) incidence, (c) cause-specific mortality, and (d) standardized mortality ratio.
The data points represent the results of many different studies conducted for many different reasons over the last 50 years. Study-level fixed effects, discussed in chapter 6 and demonstrated further in chapter 14, aid the model in explaining bias resulting from differing diagnostic criteria and study populations. The model finds no more bias associated with subnational studies than with national studies, estimating that the former are shifted in log-space by $-0.03$ on average, with a 95% uncertainty interval (UI) of $[-1.1, 1.2]$. Similarly, the model estimates that studies that did not use a neurologist are shifted in log-space by 0.02 with a UI of $[-0.3, 0.3]$. Studies that used a nonstandard definition for PD diagnosis, on the other hand, were found to be systematically biased to lower levels of prevalence than studies using a standard definition. The effect coefficient for the shift in log-space was estimated to have a mean of $-0.48$ and a UI of $[-0.7, -0.2]$.

Of the 21 regions reported in the GBD Study 2010, only 36 countries from 16 regions are represented in the systematic review of PD. GBD 2010 predicts year-age-sex estimates for all countries, even those without data. To predict out-of-sample year-age-sex estimates, explanatory covariates and fixed effects modeling provides a solution to the problem of missing epidemiological data, as developed in chapter 6. The model for PD uses the kilocalories of stimulants consumed per capita per day and national smoking prevalence as explanatory covariates. For an increase of one unit of stimulants per capita per day, the model estimates a shift in log-space of $-0.34$ with a UI of $[-0.5, -0.2]$. Similarly, for an increase of one unit in national smoking prevalence, the model predicts a shift in log-space of $-0.02$ with a UI of $[-0.05, -0.01]$.

Nonsampling variation that cannot be explained is another problem with such noisy and heterogeneous data. Chapter 6 explains how random effects can be used to estimate the systematic differences between countries within a region, regions within a superregion, and so on. Since this example contains data from Western Europe only, it is the country-level random effects that are relevant here. For example, the model estimates that prevalence in the Netherlands is above the regional mean, shifting estimates up by 20% (UI $[0, 50]$%). The model also estimates that prevalence in Great Britain is below the regional mean, shifting estimates down by 15% (UI $[0, 30]$%).

It is intuitive that there is a relationship between the different epidemiological parameters: every prevalent case was once an incident case, for example. Combining all parameters to produce internally consistent results is discussed in detail in section 7.2. Through the process of data con-
frontation discussed in the following chapters, the meta-analysis produces a best estimate and uncertainty bounds of disease prevalence, as shown in figure 1.2.

Figure 1.2. Estimates of age-specific rates of (a) prevalence, (b) incidence, (c) cause-specific mortality, and (d) standardized mortality ratio of Parkinson’s disease in Western European females in 2005.
1.2 From systematic review to metaregression

To put the descriptive epidemiological metaregression framework developed in this book in its historical context, we will now provide a brief overview and introduction to meta-analysis and systematic review.

Meta-analysis combines the results of several studies that address a set of related research hypotheses. In its simplest form, this technique identifies a common measure of interest in all studies, for which a weighted average might be the output of the meta-analysis. For example, the weighting could be related to sample sizes within the individual studies.

The history of meta-analysis often begins with the work of Karl Pearson. In 1904, the British military commissioned Pearson to evaluate the military’s typhoid inoculation campaigns. Pearson obtained data on typhoid inoculation and mortality from two studies, one from India and one from South Africa, but determined that both sample sizes were too small to permit a reliable analysis. To increase the sample size, he combined the data and thus embarked on the first meta-analysis in public health. Unfortunately, this landmark study concluded little. With such heterogeneous data and irregular results, Pearson found it problematic assigning how much weight should be attributed to different results. Despite its inauspicious beginning, meta-analysis continued to develop.

The technique of systematic review has developed extensively since Pearson’s time. The sheer number of publications every year has forced researchers to devise new ways to summarize and synthesize the torrent of data. From 1907, three years after the first meta-analysis, to 2007, the number of scientific publications has exploded. The number of abstracts compiled by the American Chemical Society has grown at 4.6% per year over that 100-year period. The number of publications compiled by the American Mathematical Society has grown at 5.9% per year. The number of publications in Compendex, a database of engineering studies, has grown at 3.9% per year. PubMed, the largest database of biomedical literature in the world, now contains more than 21 million citations. Despite this growth in publications, data are as heterogeneous and irregular as ever. This challenge certainly remains for data in descriptive epidemiology. Integrative systems modeling provides a framework to get the most information out of these disparate data.
As the number of scientific publications grew, identifying various sources to synthesize in a meta-analysis became a formidable task in its own right. This challenge led to the formalization of the process for identifying sources and the development of systematic review. The Cochrane Collaboration is a group of over 28,000 volunteers who review data from randomized control trials of health interventions. In addition to the valuable information they provide on the efficacy of a wide range of interventions, they have created a detailed handbook for conducting systematic reviews. The Cochrane Collaboration defines “systematic review” as the methodic and explicit identification, selection, appraisal, collection, and analysis of relevant research. Meta-analysis then is defined as the use of statistical techniques to combine the results of studies from a systematic review.

The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) group has also developed guidance for systematic reviews by standardizing the steps involved in a modern approach to the procedure. PRISMA divides the systematic review process into four stages: Identification, Screening, Eligibility, and Included. In the Identification stage, the reviewer finds citations for studies by searching databases like PubMed and by contacting individual researchers and institutions. The reviewer uses a specific set of keywords for the database search in order to make that search transparent and replicable. In the Screening stage, the reviewer removes duplicated and unusable data. In the Eligibility stage, the reviewer excludes articles that do not match the explicit criteria for inclusion in the study. For instance, some systematic reviews in epidemiology only include evidence from randomized control trials and exclude observational data. In the Included stage, the reviewer finalizes the studies used for the systematic review.

Meta-analyses rely critically on the systematic review procedure. Here it is convenient to follow the terminology used by the Cochrane Collaboration and PRISMA and use “meta-analysis” to refer to statistical methods for combining evidence. This provides a clear separation between systematic review and meta-analysis and also divides meta-analysis from nonstatistical approaches of “research synthesis” or “evidence synthesis,” such as combining information from qualitative studies.

As part of GBD 2010, groups of disease experts implemented this process for each disease and risk factor to be included. Because of the nature of data sources for descriptive epidemiology, the effort to capture data from unpublished sources was much more intense than in the traditional review.
of intervention studies in the Cochrane library. For some diseases, such as schistosomiasis, the vast majority of data obtained (more than 98%) came from studies that are not in the peer-reviewed literature. Regardless of the primary source of epidemiological measurements, the methodological challenge was then to take the resulting data as input to generate estimates of epidemiological parameters of interest such as incidence, prevalence, and duration.

By far the most common use of meta-analytic techniques in epidemiology is to estimate the effect size of an intervention. By pooling all studies of the intervention effect, the meta-analysis provides a more precise estimate of effect size than that found in any individual single study.

The Cochrane guidelines caution not to compare studies with very different outcome measures of effect or very different patient populations when conducting a meta-analysis. This is a subtle point and is more clearly developed in the effect-size meta-analysis realm than in the meta-analysis of descriptive epidemiological data. In fact, comparing studies with different outcome measures is at the heart of this book, which develops a method for comparing the results of descriptive epidemiological studies of disease prevalence, incidence, remission, and mortality risk that are focused on subpopulations from varying age groups, sexes, regions, and time periods.

Our framework is not without precedent, however. The next section discusses the legacy of “generic disease modeling,” upon which our integrative approach to descriptive epidemiological metaregression builds.
1.3 History of generic disease modeling

Research into disease modeling for descriptive epidemiology has been accompanied by software implementations since the 1990s. The development and refinement of these computer codes provide conveniently named milestones through the history of the approach. For example, DisMod I was software developed in the early 1990s to support analysis in the original Global Burden of Disease Study. Computing power has increased dramatically over the 20-year period in which the DisMod family of generic disease modeling software has evolved, and the aspiration of methods has expanded as well. We will now trace how the approach has evolved from a simple spreadsheet model to a robust metaregression framework.

The precursor to the first DisMod software, the Harvard Incidence-Prevalence (HIP) Model, was a spreadsheet implemented in Lotus 123. This model took as input a set of instantaneous incidence, remission, and excess mortality rates for 5 age groups and produced estimates of disease prevalence and duration. The model involved constructing a life table to simulate a cohort exposed to a set of age-specific incidence, remission, case fatality, and background mortality hazards. At each year in the life table, the model simulated a simple 3-compartment model to provide estimates of the number susceptible, the number of cases, and the number of deaths to input into the life table for the next year. It was used primarily for three purposes: to find prevalence for conditions where incidence is known and reasonable assumptions about remission and excess mortality can be made; to find attributable deaths that were not directly coded to a specific cause; and to find incidence for conditions where prevalence is known. The third use required an interactive procedure in the HIP Model, since the input incidence was unknown.

As is often the case in science, a very similar approach had been developed previously by researchers at the International Institute for Applied Systems Analysis in Austria in the 1970s. This work was part of a broad program to develop a generic Healthcare System Model to improve management and planning in the health sector. One component of this model was a computer program to estimate prevalence from incidence. That program evolved a population exposed to age-specific incidences of disease and death through time. Although it was designed specifically for terminal illness, it is similar to the DisMod line of models in many ways. It was applied to estimate the prevalence of malignant neoplasm in Austria, France, and
Over the course of the first Global Burden of Disease Study, the HIP Model evolved in DisMod I. This was formalized as a 4-compartment model and a corresponding system of differential equations. As in the HIP Model, the input to DisMod I consisted of instantaneous rates for incidence, remission, and excess mortality, now specified for 9 age groups. In addition, DisMod I was also used to estimate the average duration of disabling sequelae as a function of age. DisMod I was used iteratively by analysts working on GBD 1990 to identify a solution that matched the available data on prevalence, incidence, remission, excess mortality, and cause-specific mortality. DisMod I was used to address multiple challenges: mapping from incidence data to prevalence and vice versa and assessing the consistency among incidence, prevalence, and cause-specific mortality.

DisMod II moved from forward simulation into the realm of optimization. It provided more control over inputs, as well as a graphical user interface and comprehensive user manual, making it more widely usable than previous iterations. In addition to accepting input consisting of instantaneous rates for incidence, remission, and excess mortality, DisMod II was also capable of using age-specific prevalence and cause-specific mortality rates, as well as incidence as a population rate and duration when it is short (less than 1 year). It also provided an algorithmic method for data confrontation wherein the downhill simplex method was used to minimize the weighted difference between the inputs and the output predictions. Although DisMod II included the optimization, it was not framed as a statistical likelihood estimation and thus did not generate statistical uncertainty intervals. To put it another way, it did not include a measurement model, only the model relating true population rates.

The World Health Organization (WHO) distributed the DisMod II software without cost, and thus, the generic disease modeling approach has been used widely in burden of disease studies over the last 10 years. These studies adopted the methodology of the global study but aimed to assess burden at a level of detail more relevant for national policymakers. At least 37 countries have undertaken national or subnational burden of disease studies, including Mexico, Chile, Colombia, and Mauritius.

Despite its wide application, DisMod II has been criticized. One methodological concern that emerged from extensive application of the model centered on the difficulty in producing consistent estimates that exhibited face validity—for example, age patterns that increased monotonically as a func-
tion of age. Despite strongly held prior beliefs on the part of domain experts, it was not uncommon for the prediction to show oscillations as a function of age, due to the contortions to which DisMod II would subject rates in order to produce consistent estimates as close to the single-rate-type input estimates as possible.

Another important challenge in the DisMod II work flow was the production of single best estimates for at least three independent rates to be used as input. Systematic review often finds multiple measurements of an age-specific rate, and only one could be the input to DisMod II. Transforming a large collection of measured values, often for incommensurate age intervals, to a single best estimate of disease prevalence was a difficult analytic challenge that was a necessary preprocessing step to do meta-analysis with DisMod II.

A third challenge with DisMod II was in producing robust estimates of parameter uncertainty. Although the system included a method to propagate uncertainty in the input parameters through to the output estimates, this was laborious and rarely used in practice.

Finally, although DisMod II excelled in providing consistent estimates from inconsistent estimates of several disease parameters for a single place and time, it was laborious for the data analyst to produce comparable estimates for a variety of different places and times. In the GBD Study 2010, there are 21 geographic regions to produce estimates for, at three different points in time, for males and females. Even an analysis that is trivial for one region/time/sex becomes burdensome when it must be replicated 126 times.

For GBD 2010, we completely redeveloped the method, continuing the trend toward including more formal inferential techniques in the estimation process. The broad principle behind this approach is what we call integrative systems modeling (ISM) and can be characterized in two parts: a system dynamics model of process and a statistical model of data, considered together, so that instead of doing forward simulation, as is traditionally the case in system dynamics modeling, the model is used to solve an inverse problem. This method is emerging as a powerful approach for developing models that integrate all available data sources. On top of the compartmental model initially conceived for the HIP Model, we have layered an age-standardizing, negative-binomial, mixed-effects spline model, which is fitted directly to the data extracted in systematic review using Bayesian methods.
This meta-regression technique has been implemented in a free/libre open-source software package called DisMod-MR. The details of the approach constitute the bulk of the first part of this book. The second part is dedicated to a series of example applications, demonstrating concretely the features developed in Part I.
Part I

Theory
Chapter 2

Statistical models for rates, ratios, and durations

Abraham D. Flaxman

A central decision when modeling sparse, noisy epidemiological data with Bayesian methods is the choice of the data likelihood function. That is why the framework begins with a thorough discussion of the alternative models for the meta-analysis of rates, ratios, and durations.

A useful thought experiment to guide the development of a metaregression technique is to consider how the model would proceed if, for each and every study identified in systematic review, complete microdata were available. Of course, it is unusual that microdata are available for even one study from the systematic review. However, if all the microdata were available, say for all the prevalence studies conducted on schizophrenia (an example I will return to in the next section), modeling could proceed through standard techniques for analyzing binary data, such as logistic or probit regression, with fixed effects to explain some of the nonsampling variation, such as differing diagnostic criteria, and random effects to model the additional nonsampling variation, such as inherent differences between populations (if they exist).

Viewed in this light, the task of a metaregression model is to produce the results that would be obtained from an analysis of all the microdata, if they were available. The approach that will be developed below decomposes into three parts: the epidemiological rate model, which captures the sampling error in systematic review data; the age-interval model, which addresses the heterogeneity of age groups reported in the literature; and the covariate
model, which models the nonsampling variation between different sources of data through fixed and random effects.

The key to connecting the data for different epidemiological parameters, such as incidence and prevalence, is the systems dynamics model presented in chapter 7. This model of process describes how the models of data are related to each other. I will return to the model of process once the model of data is fully developed. The model of data is a statistical model that has its core features defined by its likelihood function. By likelihood function, I mean a probability density function that assigns a value for the likelihood of every possible observed value for any setting of the model parameters.

2.1 A motivating example: Schizophrenia prevalence

An example will make this clearer, so I turn now to the meta-analysis of population prevalence of schizophrenia in adult males. Strictly speaking, prevalence is a ratio, although in the literature the term “prevalence rate” is often used to mean prevalence ratio. The prevalence of a condition in a population says something about the stocks in the stock-and-flow model from chapter 7 for a specific time period and age group, prevalence is the ratio of individuals with the condition to all individuals in the population.

The forest plot in figure 2.1 shows the results of combining 16 studies using 7 different data models. As the figure demonstrates, the choice of data model can have a substantial effect on the estimated median, as well as on the uncertainty. The models I display produce point estimates ranging from 1.2 to 4.0 per 1000 person-years and uncertainty intervals with widths ranging from 0.1 to 2.9. When analyzing sparse and noisy data, the choice of the data model matters.

In what follows, I will develop a collection of data models, starting with the simplest and then increasing the complexity, while identifying the benefits and drawbacks of each. The models to come, in order, are

- the binomial model,
- the beta-binomial model,
- the Poisson model,
2.2  BINOMIAL MODEL

Figure 2.1. Forest plot summarizing 7 alternative models for meta-analysis of adult male schizophrenia prevalence at the population level. The median estimates range from 0.0012 to 0.004 per person-years, and the width of the 95% uncertainty interval ranges from 0.1 to 2.9.

- the negative-binomial model,
- three variants of the normal model,
- the lower-bound data model.

2.2  Binomial model

Conceptually, the simplest model I consider for epidemiological data is built from the binomial random variable. Random variable $X$ is binomially di-
tributed if it has probability distribution

$$P[X = k \mid n, \pi] = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

for some $\pi$. I have used Greek to emphasize that $\pi$ is a model parameter, while $n$ and $k$ are data.

Although this equation may appear opaque, the intuition behind it is simple: $n$ individuals were tested for a disease, and $k$ tested positive. The formula then follows from the assumption that each individual tested positive with probability $\pi$, and that if I know $\pi$, then knowing about the test results of any subset of individuals gives me no information about the test results for the others (these events are “independent”).

The intuitive description in the previous paragraph is particularly relevant to a study that measures the prevalence of a disease in a population. As mentioned above, prevalence is a ratio, not a rate. It is a unitless metric, often expressed as a fraction or a percentage. Incidence, remission, and mortality rates, on the other hand, are rates, measured per unit time. For example, incidence is often expressed in the units of “per person-year” or “per 1000 person-years.” Nonetheless, the binomial distribution can be the basis of a statistical model for rates as well as for prevalence. I will argue that it is not a very good model, and the fact that, in its intuitive description, it may have the wrong units is only one of its shortcomings. It is instructive to begin with simple models and make them more complicated until they are just complicated enough.

The binomial distribution inspires a computationally tractable data model for an observed population prevalence rate of $r$ in a sample population of size $n$:

$$p(r \mid \pi, n) \propto \pi^{\left\lfloor rn \right\rfloor} (1 - \pi)^{\left\lceil (1-r)n \right\rceil}.$$  

Here $p(\cdot)$ denotes a probability density function, $\left\lfloor \cdot \right\rfloor$ is the “floor” operator, which rounds real numbers down to the largest integer less than or equal to the operand, and $\left\lceil \cdot \right\rceil$ is the “ceiling” operator, which rounds up.

Note that it is not necessary to include the normalization term $\binom{n}{\left\lfloor np \right\rfloor}$, because this does not depend on the model parameter $\pi$. This constant of proportionality is necessary to make this data model truly a probability density function for any $\pi$ and $n$. But I will never need to know this constant, and I use the “proportional to” symbol $\propto$ instead of equality to emphasize this fact.
2.2. BINOMIAL MODEL

In terms of the thought experiment from the introduction, this model is equivalent to an analysis of all available microdata by fully pooling all individual measurements from all studies. It simply uses the sample population from each study together with the rate to find the number of positive observations. In the parlance of meta-analysis, this is a “fixed-effect meta-analysis,” because the rate is modeled as fixed across all populations.

The funnel plot in figure 2.2 shows the predictive distribution of this rate model for \( \pi = 0.004 \). The handful of square markers show the observed rate \( (r) \) and study size \( (n) \) from 16 studies of schizophrenia prevalence. The more plentiful circular markers are intended to give some idea of the shape of the funnel predicted by the binomial rate model. This figure shows the potential problem with this approach: the data gathered by systematic review are often much more dispersed than this distribution predicts. The binomial model row of the forest plot in figure 2.1 shows the implications of this problem in the case of schizophrenia prevalence: two large studies are responsible for pulling the estimate below most of the observations, while the pooled uncertainty interval is so small that it does not overlap the uncertainty of most of the data points.

Figure 2.2. Funnel plot showing predictive distribution for the binomial rate model with \( \pi = 0.004 \) (samples from this distribution are marked by circles), with data from systematic review for adult male schizophrenia prevalence overlaid for comparison (observations marked by squares).

There are two clear problems with this model—biased estimates and unreasonably high confidence when modeling noisy data. The model ap-
appears biased because many measurements are larger than the upper limit of the uncertainty interval, while none are smaller. The uncertainty interval appeared too small because it does not account sufficiently for noise in the measurement of \( r \). If a study of 50,000 people from subpopulation A finds prevalence of 2 per 1000 and a study of the same size in subpopulation B finds 6 per 1000, then the binomial model predicts that a third study conducted in subpopulation C will have prevalence of 4 per 1000, with an uncertainty interval of [3, 5] per 1000 (I use the 95% highest posterior density [HPD] interval as the uncertainty). I have no problem with the point estimate; picking the mean of the two populations seems just right. But the uncertainty interval lacks face validity. It would be much more reasonable to have an uncertainty interval as large as [1, 7] instead of one as small as this.

Another way to quantify the mismatch between the binomial rate model and the observed data is through the posterior predictive check, an in-sample goodness-of-fit test that graphically compares the observed data to the posterior predictive distribution, which is to say the model’s prediction for what the data should be, after it has been fitted to the data. Figure 2.3 shows 1000 draws from the posterior predictive distribution for each of the data observations, together with the observation itself. The model predictions are compressed, showing a cloud of predicted points that often does not include the observation. Trusting the results of such a model leads to inappropriately high certainty about the nature of these noisy data.

### 2.3 Beta-binomial model

A theoretically appealing extension to the binomial model (which also is not sufficient for my purposes) is the beta-binomial model. I will develop it in this section to motivate the following sections.

Formally, a beta-binomial random variable \( X \) has the following probability distribution:

\[
P[X = k \mid n, \alpha, \beta] = \int_{\pi=0}^{1} p(\pi \mid \alpha, \beta) \binom{n}{k} \pi^k (1-\pi)^{n-k} d\pi,
\]

\[
p(\pi \mid \alpha, \beta) \propto \pi^{\alpha-1} (1-\pi)^{\beta-1}.
\]

The intuition behind this model is simpler than the equation, however. As in the binomial model, each individual tests positive for the condition
2.3. BETA-BINOMIAL MODEL

Figure 2.3. Posterior predictive check for binomial model fitted to adult male schizophrenia data. Circles show 1000 draws from the posterior distribution of the binomial model, and squares show the observed data. The uncertainty interval marked by error bars around each square shows the sampling error for each observation, based on the sample size alone. More than half of the data observations fall below the posterior predictive distribution samples, indicating that the model is biased and not capturing the heterogeneity observed in the data.

independently with a probability \( \pi \), but now \( \pi \) itself is a random variable, distributed according to a beta distribution with parameters \( \alpha \) and \( \beta \). The beta distribution is given by

\[
p(\pi | \alpha, \beta) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1}
\]

and has a high degree of flexibility. It always takes values between 0 and 1, making it an appropriate distribution for a probability. Figure 2.4 shows the probability density of the beta distribution for several combinations of \( \alpha \) and \( \beta \).

The beta-binomial distribution inspires the following data model for an observed rate of \( r \) in a population of size \( n \):

\[
p(r | \alpha, \beta, n) \propto \int_{\pi=0}^{1} \pi^{\alpha-1}(1-\pi)^{\beta-1}\pi^{\lfloor rn \rfloor}(1-\pi)^{\lceil (1-r)n \rceil} \, d\pi.
\]

This model extends the binomial model in a way analogous to a random-effects model in traditional meta-analysis (19) (or in linear regression (21)). By introducing additional dimensions into the parameter space, the model is able to capture the dispersion beyond the binomial model that I have observed empirically in funnel plots of real data. Figure 2.5 shows the
Figure 2.4. Probability density for the beta distribution for a range of $\alpha$ and $\beta$ values. The dashed vertical line shows the expected value, which is one-half for all distributions in panel (a) and one-fourth for all in panel (b).

beta-binomial funnel plot, as well as the posterior predictive check for this model on the same data as used in figure 2.3.

This model addresses the theoretical shortcoming raised in the previous section: if studies of 50,000 people show prevalences of 2 and 6 per 1000, then the posterior distribution of the beta-binomial model has mean 4 with uncertainty interval [1, 8], which seems quite reasonable.

The great shortcoming of the beta-binomial model is computational. There is no closed-form solution to the integral in the probability density for the beta-binomial model. Evaluating it requires introducing a latent variable for each of the data points in the likelihood. This simply takes too long to compute for the numerical algorithms and computational infrastructure available in 2012.

2.4 Poisson model

There are two traditional approximations to the binomial distribution, depending on how large $k$ is in relation to $n$. When $k/n$ is large, the normal distribution is used, and when $k/n$ is small, the binomial is similar to the Poisson distribution.

Since I expect disease prevalence to usually fall in a “small $k/n$” setting, I will not develop the normal model in detail now, although in section 2.6
2.4. POISSON MODEL

![Funnel plots and posterior predictive check for the beta-binomial model. This model captures the heterogeneity in the observed data more faithfully than the binomial model from the previous section. However, the estimation procedure requires the introduction of a latent parameter for each data observation, and integrating out these latent variables is too computationally demanding to be feasible in many current applications.](image)

I will develop a model based on monotonic transformations of the normal distribution that include a normal model as a special case.

The Poisson distribution is given by the equation

\[ P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}, \]

and it can be understood intuitively as the number of times a “memoryless” event occurs in a unit time period. Setting \( \lambda = \pi n \) produces an approxi-
formation to the binomial distribution that is quite accurate for large \( n \) and small \( k \).

My Poisson rate model, which is defined from the Poisson distribution in a manner analogous to the way the binomial distribution was converted to a binomial rate model above, is the following:

\[
p(r \mid \pi, n) \propto (\pi n)^{rn} e^{-\pi n}.
\]

It is also subject to all of the concerns raised about the binomial model. When modeling rates with nonsampling variation at the level typically found in systematic review, it will produce inappropriately low estimates of uncertainty.

There is one key benefit to this model compared to the binomial and beta-binomial models, however. The Poisson model assigns a nonzero likelihood to rates of more than 1. Although prevalence is always less than 1, it is theoretically possible to have incidence rates more than 1 (per person-year), and remission rates are often more than 1. For prevalence, which, as discussed above, is actually a unitless ratio of cases to population size, having nonzero probability of values greater than 1 is incorrect. But for incidence, remission, and excess mortality, which are measured per unit time, having a model with a natural interpretation in the same units is appealing.

### 2.5 Negative-binomial model

Another benefit from the count model approach is to be found in the Poisson model’s overdispersed cousin. This distribution is called the negative-binomial distribution (named after the formula that proves it does indeed sum to 1):

\[
P[X = k \mid \pi, \delta] = \frac{\Gamma(k + \delta)}{\Gamma(\delta)k!} \left( \frac{\delta}{\pi + \delta} \right)^\delta \left( \frac{\pi}{\pi + \delta} \right)^k.
\]

Unlike the beta-binomial distribution, it can be approximated accurately without numerical integration.

However, this closed form obscures the intuition behind the negative-binomial distribution, which is quite similar to the intuition behind the beta-binomial distribution (but less clear from its name). Through a bit of algebra, the negative-binomial distribution can be represented as a hierarchical model where the observed data come from a Poisson distribution,
and the parameter of the Poisson distribution is itself a random variable
that comes from a gamma distribution:

$$X \mid \lambda \sim \text{Poisson}(\lambda),$$
$$\lambda \sim \text{Gamma}(\pi, \delta).$$

Here the gamma distribution is defined by

$$p(\lambda \mid \pi, \delta) \propto \lambda^{\delta-1} \exp \left(-\lambda \delta / \pi\right).$$

The identity

$$\frac{\Gamma(k + \delta)}{\Gamma(\delta)k!} \left(\frac{\delta}{\pi + \delta}\right)^\delta \left(\frac{\pi}{\pi + \delta}\right)^k = C_{\pi, \delta} \int_0^\infty \frac{e^{-\lambda} \lambda^k}{k!} \lambda^{\delta-1} e^{-\lambda \delta / \pi} \, d\lambda$$

for an appropriate constant $C_{\pi, \delta}$ verifies this interpretation.

Through this lens, the negative-binomial model can be seen as a natural
adaptation of the traditional random effects model in linear regression to
the Poisson case, where each observation comes from a different Poisson
model and the Poisson parameters of these models are all drawn from a
common gamma distribution. Thus, a rate model based on this distribu-
tion provides benefits in handling nonsampling variation similar to those
demonstrated for the beta-binomial distribution above but in a formulation
that is much less demanding computationally. The negative-binomial rate
model for observing a rate of $r$ in a population of size $n$ is

$$p(r \mid \pi, \delta, n) \propto \frac{\Gamma(\lfloor r n \rfloor + \delta)}{\Gamma(\delta)} \left(\frac{\delta}{\pi + \delta}\right)^\delta \left(\frac{\pi}{\pi + \delta}\right)^{\lfloor r n \rfloor}.$$ 

Figure 2.6 shows funnel plots for two levels of overdispersion, as well as
the posterior predictive distribution for the negative-binomial model.

2.6 Transformed normal models

Some epidemiological data are not related to count data at all. Duration
studies and studies measuring the relative risk of mortality are two that
come up frequently in systematic review. For duration data, a normal
model is sufficient, and a lognormal model proves to be appropriate for
Figure 2.6. Funnel plots and posterior predictive check for the negative-binomial model. This model captures heterogeneity in observed data using an “overdispersion” parameter $\delta$ and can be interpreted as a hierarchical model, where each observation is drawn from a Poisson distribution that has its parameter drawn from a gamma distribution. When $\delta$ is very large, the negative-binomial model is equivalent to the Poisson model. The posterior predictive check shows that the uncertainty interval in the posterior predictions contains all the observed data, indicating that this model is sufficiently flexible to represent the observed heterogeneity.

modeling relative risk data, which can be thought of as a ratio of count variables.

Transformed normal models have also been used for mortality rates in the past \cite{22,23,24} and are worthy of continued consideration for modeling the incidence, prevalence, remission, and mortality as an alternative to the negative-binomial model.
In this section, I will develop a general transformed normal model and compare it to the negative-binomial model. The adjective “transformed” refers to a function that I will keep quite general for now, only requiring it to be increasing and differentiable; that is, for any \( x > y \) the transformation \( f \) must have \( f(x) > f(y) \) and \( f'(x) \) must be defined. Then the transformed normal model will be derived from the normal distribution, defined by the probability density

\[
p(x \mid \pi, \sigma) \propto \frac{1}{\sigma} \exp \left\{ -\frac{(x - \pi)^2}{2\sigma^2} \right\}.
\]

For any increasing, differentiable function \( f \), this distribution can be converted to an \( f \)-transformed normal model with probability density

\[
p(r \mid \pi, \sigma, f, s) \propto \exp \left\{ -\frac{[f(r) - f(\pi)]^2}{2\left[(sf'(r))^2 + \sigma^2\right]} \right\},
\]

where \( s \) is the standard error of the rate \( r \), which is more convenient than the effective sample size \( n \) in this case. The denominator of the exponent deserves some additional discussion. For the identity function \( f(x) = x \), the derivative \( f'(x) = 1 \), and the denominator simplifies to \( 2(s^2 + \sigma^2) \), a familiar “inverse variance” weighting, where \( \sigma \) is a random effect to account for overdispersion. When \( f \) is a more complicated function, the term \( sf'(r) \) approximates the standard error of the transformed value \( f(r) \). Although more sophisticated approximations are possible, experience dictates that the nonsampling variation (parametrized by \( \sigma \)) is always larger than the chance variation, so a simple approximation of the chance variation is sufficient.

Some common transformations of \( f \) used in related work yield the log-normal model \( f(x) = \log x \), the logit model \( f(x) = \logit(x) \), and the probit model \( f(x) = \text{probit}(x) \). All these approaches have a significant drawback, however. The transformation is not defined for \( x = 0 \), so these models cannot use data showing rates of 0. There are two common methods to fix this: dropping all 0s and adding a small offset. Dropping measurements of 0 is clearly problematic, as it leads to systematic bias in the data that remain and produces estimates larger than the truth. This is especially problematic for high-quality studies that focus on the age pattern of a disease, where it is quite reasonable for some age groups to have 0 cases observed. The effect of dropping 0s is to overestimate the rates in these age groups.
CHAPTER 2. STATISTICAL MODELS FOR RATES

Adding a small offset, such as 0.5, is an alternative solution, and indeed, this is the approach taken for cause-specific mortality estimation in a similar approach to mortality modeling. The selection of the offset can seem ad hoc, however.

Within the framework of the transformed normal model, there is room to put the solution on firm theoretical foundations. For example, by taking \( f_\zeta(x) = \log(x + \zeta) \), I obtain the “offset log-transformed model,” which does allow rates of 0, simply by taking a positive value for \( \zeta \). This model will not be used extensively in the example application to come later in this book, but it seems like a promising approach. It is particularly appealing in the way it decomposes the sampling variation into an additive error \( \zeta \) and a multiplicative error \( \sigma \), and I expect that it will prove useful in the future. For completeness, here is the probability density for the offset log-transformed model:

\[
p(r | \pi, \sigma, \zeta, s) \propto \exp \left\{ -\frac{[\log(r + \zeta) - \log(\pi + \zeta)]^2}{2 \left( \frac{s}{r + \zeta} \right)^2 + \sigma^2} \right\}.
\]

Figure 2.7 shows funnel plots for two levels of overdispersion, as well as the posterior predictive distribution for the offset lognormal model.

2.7 Lower-bound data model

Cause-specific mortality rates (CSMRs) are a special case among the epidemiological rates available for integrative systems modeling of disease in a population. These data come from careful processing of vital registration system outputs, from verbal autopsy studies, and from some other sources. But when I introduce the compartmental model of disease in a population in chapter 7 there will be an important distinction about where CSMR data fit in. Unlike incidence, prevalence, and remission data, they do not correspond directly to any rate in the compartmental model. This is because of the operational requirement in vital registration systems that each death have a single underlying cause. The excess mortality rate in the system dynamics model from chapter 7 is not entirely compatible with this idea.
2.7. LOWER-BOUND DATA MODEL

Figure 2.7. Funnel plots and posterior predictive check for the offset log-normal model. This model captures heterogeneity in observed data using a dispersion parameter $\sigma$ and also includes an offset parameter $\zeta$ which interpolated between log-normal and normal models.

In theory, the compartmental model could be extended to include CSMR data explicitly simply by splitting the excess-mortality hazard $h_f$ flowing out of the with-condition compartment $C$ into two parts. These parts, $h_{f'}$ and $h_{f''}$, would sum to $h_f$. The quantity $h_{f'}$ would denote the portion of excess mortality caused directly by the disease, so any individuals that exit compartment $C$ via the flow with hazard $h_{f'}$ would have this condition listed as the underlying cause of death on their death certificate. The quantity $h_{f''} = h_f - h_{f'}$ would then denote the “excess excess mortality,” which is to say the elevated mortality among individuals dying with the condition but not of the condition.
Proceeding down this path promises to be extremely confusing! It is also a challenge because the sparse and noisy data available have never been sufficient to separately estimate \( h_f' \) and \( h_f'' \) with much accuracy. When the model is more flexible than the data, it is hard to fit and the results are hard to interpret.

This motivates the alternative approach that I have taken, which implicitly separates excess mortality into \( h_f' \) and \( h_f'' \) but does not try to explicitly represent both in the model. It is a “lower-bound” likelihood, which contributes nothing to the likelihood as long as the observation is below the prediction, and uses the Poisson rate model when the observation is above the predicted level:

\[
p(r \mid \pi, n) \propto \begin{cases} (\pi n)^{\lfloor rn \rfloor} e^{-\pi n}, & \text{if } \pi < r; \\ 0, & \text{otherwise.} \end{cases}
\]

### 2.8 Quantification of uncertainty

The quantification of uncertainty in this metaregression challenge is worthy of special attention. The binomial, beta-binomial, Poisson, and negative-binomial models developed above all rely on a quantification of uncertainty in terms of persons or person-years, denoted by \( n \). This is a stylized notion, however, based on a simple model of the data generation process that uses a simple random sample. In systematic review, it is common to find more complex survey designs, and more sophisticated quantification of uncertainty is often reported. I call the corresponding \( n \) the “effective sample size,” because it denotes the size that the sample would be if an identically powered study did use simple random sampling.

The transformed normal models above require a quantification of uncertainty in terms of standard error, denoted by \( s \). Many studies collected in systematic review report this value directly, but many others do not.

It is useful to have a simple set of conversions to translate between the \( n \) needed for the count models, the standard error needed for the transformed normal models, and the often-reported 95% confidence interval, which does not appear directly in any of the rate models above. The approximate relationships between these quantities are standard, and developing more precise transformations is not justified due to the large amount of nonsampling variation in systematic review data.
To represent a standard error $s$ in terms of a 95% confidence interval $(a, b)$, I have used the normal approximation

$$s = \frac{b - a}{2 \cdot 1.96}.$$

To represent an effective sample size $n$ in terms of a standard error $s$ and an observed rate $r$, I have two options. For prevalence data, where the standard error is for a ratio and hence constrained to be between 0 and 1, I have used the binomial approximation

$$n = r(1 - r) \frac{s^2}{s^2}.$$

For other epidemiological rates, such as incidence and remission rates, where the standard error is for a rate that is nonnegative but could potentially be larger than 1, I prefer the Poisson approximation

$$n = \frac{r}{s^2}.$$

Some studies from systematic review report point estimates for age-specific rates but quantification of uncertainty only at coarser levels of aggregation, for example, only the sample size of the entire study. In these cases, I recommend a rough approximation that splits the $n$ for the entire population among the subpopulations proportionally to the population age structure.

Surprising as it may be, some studies from systematic review do not report quantification of uncertainty at all. It may be acceptable to exclude these studies, and this exclusion criterion should ideally be articulated at the beginning of the systematic review process. Sometimes data are so sparse that it is not feasible to exclude studies that do not quantify uncertainty, however. In this case, I recommend imputing the missing uncertainty interval (UI) values by taking them to have UI width equal to the 90th percentile value of the UI widths from the observations that do quantify uncertainty.

2.9 Comparison

Making a quantitative comparison of the models at this point is challenging. A simulation study depends critically on the distribution used to simulate
the data sets; so choosing a simulation distribution which matches the assumptions of any model will yield results that make the chosen model look superior. A comparison based on out-of-sample predictive accuracy would be preferable, but disease data are so sparse and noisy that such a comparison could be meaningless, especially without including the adjustments for covariate effects and age integration, which are developed in the next two chapters. A handful of diseases have data homogeneous enough in age and geography to attempt a comparison of the models, but these are necessarily special cases, and extending such findings to other settings must be done with caution. On the other hand, something is better than nothing.

To this end, I compared all models using a holdout cross-validation approach. I used the Markov Chain Monte-Carlo approach, described in chapter 8, to generate 1000 samples of the model parameters from the joint posterior distribution for each model, using random subsets of the data on schizophrenia prevalence and on epilepsy prevalence collected from systematic review in the model likelihood. The schizophrenia data set is familiar from earlier in this chapter, and contains only 16 observations. The epilepsy data set contains many more rows of data, with 1719 observed prevalence values, and the level appears to vary little as a function of age, time, or geography. I included each observation in the subset to fit independently with probability 0.75, yielding a subset with 8 rows expected for schizophrenia and around 1300 observations for epilepsy. Then I fitted each model with this subset and used the fit to predict values for the observations that were not included in the subset. For 100 replicates, I measured the bias, median absolute error, coverage probability, and computation time for each model. The results are shown in table 2.1.

For the schizophrenia data, the beta-binomial model has the lowest MAE, but takes the longest to run, while the negative-binomial, normal, offset lognormal and lognormal models all have slightly higher MAE, and similarly low bias and accurate calibration. The binomial and poisson models have substantially more bias, as well as higher MAE and lower PC.

For the epilepsy data, the beta-binomial model has coverage probability 10 times higher than the binomial and Poisson models, which also minimize median absolute error, but takes 1000 times longer to compute and is still far from the target probability of coverage of 95%. The lognormal model has the superior balance of MAE and PC, but cannot cope with data measurements of zero prevalence. All of these four models have substantial bias as well. In addition to being theoretically justified, the negative-binomial
2.9. COMPARISON

Table 2.1. Mean results of holdout cross-validation of rate models for 100 replicates for out-of-sample prediction, ordered by increasing median absolute error (MAE).

Schizophrenia data:

<table>
<thead>
<tr>
<th>Rate Model</th>
<th>Bias (pp)</th>
<th>MAE (pp)</th>
<th>PC (%)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta-binomial</td>
<td>0.08</td>
<td>0.08</td>
<td>94.6</td>
<td>213</td>
</tr>
<tr>
<td>Negative-binomial</td>
<td>0.00</td>
<td>0.16</td>
<td>96.5</td>
<td>76</td>
</tr>
<tr>
<td>Normal</td>
<td>0.01</td>
<td>0.17</td>
<td>94.3</td>
<td>63</td>
</tr>
<tr>
<td>Offset lognormal</td>
<td>0.01</td>
<td>0.17</td>
<td>92.7</td>
<td>84</td>
</tr>
<tr>
<td>Lognormal</td>
<td>-0.06</td>
<td>0.19</td>
<td>98.1</td>
<td>73</td>
</tr>
<tr>
<td>Binomial</td>
<td>0.28</td>
<td>0.25</td>
<td>11.4</td>
<td>49</td>
</tr>
<tr>
<td>Poisson</td>
<td>0.28</td>
<td>0.25</td>
<td>11.4</td>
<td>49</td>
</tr>
</tbody>
</table>

Epilepsy data:

<table>
<thead>
<tr>
<th>Rate Model</th>
<th>Bias (pp)</th>
<th>MAE (pp)</th>
<th>PC (%)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta-binomial</td>
<td>0.33</td>
<td>0.26</td>
<td>29.8</td>
<td>182299</td>
</tr>
<tr>
<td>Binomial</td>
<td>0.38</td>
<td>0.26</td>
<td>3.2</td>
<td>101</td>
</tr>
<tr>
<td>Poisson</td>
<td>0.38</td>
<td>0.26</td>
<td>3.2</td>
<td>83</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.25</td>
<td>0.27</td>
<td>88.5</td>
<td>124</td>
</tr>
<tr>
<td>Negative-binomial</td>
<td>0.09</td>
<td>0.36</td>
<td>88.5</td>
<td>190</td>
</tr>
<tr>
<td>Normal</td>
<td>0.00</td>
<td>0.42</td>
<td>95.1</td>
<td>105</td>
</tr>
<tr>
<td>Offset lognormal</td>
<td>0.00</td>
<td>0.42</td>
<td>95.1</td>
<td>130</td>
</tr>
</tbody>
</table>

Note: Bias is the mean of observed minus predicted measured in percentage points (pp), median absolute error (MAE) is the median of the absolute difference between observed and predicted, probability of coverage (PC) is the fraction of observed falling within 95% uncertainty interval of prediction, and time is the computation time.
model has a balance of low MAE and high PC, with much lower bias. The Normal and offset lognormal models have even lower bias and are perfectly calibrated, but have the highest MAE.

2.10 Summary and future work

This chapter has developed seven alternative rate models, all with benefits and drawbacks. The binomial model is simple and theoretically appealing but does not handle nonsampling variation, producing overconfident estimates in the face of noisy data. The beta-binomial model deals with overdispersion through a theoretically appealing extension to the binomial model, but it is too computationally demanding to use in my applications. The Poisson model is a close approximation to the binomial model and has all the drawbacks except that it can handle rates greater than 1, which is important for modeling remission rates. So it is the negative-binomial model, which extends the Poisson model analogously to the way the beta-binomial model extends the binomial model, that I prefer on theoretical grounds. This is the model that I have used for most of the applications to follow, where it is applied to represent data for incidence, prevalence, remission, excess mortality, and cause-specific mortality. It is not as amenable to analysis as I would like, however, and sometimes benefits from weakly informative priors on the overdispersion parameter, an undesirable feature that slows down analysis by requiring sensitivity analysis. Transformed normal models are a promising alternative approach, and I have used the normal model for duration data in some of the following examples, as well as the lognormal model for standardized mortality rate data and relative mortality risk data. The offset log-transformed model seems particularly promising as an alternative to the negative-binomial model, and understanding its statistical and computational characteristics is a promising direction for future research. Like the negative-binomial model, however, the convergence of these transformed normal models also benefits from weakly informative priors, a topic that deserves additional attention in the future.
Chapter 3

Age pattern models

Abraham D. Flaxman

The rate models of data in the previous chapter need several extensions to be truly useful in descriptive epidemiological metaregression. The most important is representing the differences in rates as a function of age. In this chapter, I develop the mathematical and statistical theory behind a model for age specificity in prevalence rates as well as other epidemiological hazard functions, such as incidence, remission, without-condition mortality, and excess-mortality hazards.

Figure 3.1 shows age-specific all-cause mortality rates for 5-year age groups. These mortality estimates are for females in Southern sub-Saharan Africa in 1990. A striking feature of this plot is the range of variation in mortality levels as a function of age. They vary by 850-fold between the minimum in the 10- to 14-year-olds and the maximum at the oldest ages. Epidemiological rates also vary among regions, times, and sexes. A figure like figure 3.1 for the high-income Asia Pacific region in 1990 would look very different, as would the Southern sub-Saharan Africa region in 2010. However, systematic variation as a function of age is often the largest source of variation by orders of magnitude, and furthermore, this variation is often distinctly nonlinear.

The approach that I have taken for modeling age-specific hazards draws on the mathematical theory of spline interpolation and on the statistical theory of penalized spline regression. It is developed in full detail in this chapter.
3.1 Definition of spline models

For my purposes, a spline model can be any piecewise polynomial function. Often I will require this function to be continuous, but not always. This is a departure from the conventions of statistical spline modeling, which focuses on continuous and continuously differentiable splines.\(^{(25,26)}\)

I represent a spline model for an age-specific hazard \(h(a)\) by a set of knots \(a_1, \ldots, a_K\) and a set of piecewise polynomial basis functions \(\{p_1, \ldots, p_{K'}\}\). Each knot has a corresponding basis function, and for higher-order splines, there may be additional basis functions as well, so \(K \leq K'\). The model then has \(K'\) parameters, \(\gamma_1, \ldots, \gamma_{K'}\), and takes the form

\[
h(a) = \sum_{k=1}^{K'} \gamma_k p_k(a).
\]

The mathematical definition of the model is straightforward, but the detail of selecting the piecewise polynomials remains to be developed. This is where the spirit of spline modeling resides. The knots \(a_1, \ldots, a_K\) partition the age range into intervals. If I make each piecewise polynomial \(p_k(a)\) equal to 1 on its interval (i.e., when \(a_k \leq a < a_{k+1}\)) and 0 otherwise, this yields a piecewise constant spline model. This is an important specialization, the simplest of my spline models. Using the notation \(1[a_k \leq a < a_{k+1}]\) to

---

**Figure 3.1.** All-cause mortality for females in the Southern sub-Saharan Africa region in 1990 as a function of age shows the range of variation in age-specific rates. All-cause mortality is as low as 6 per 10,000 PY at age 10 but rises above 5000 per 10,000 PY at age 100.
3.1. DEFINITION OF SPLINE MODELS

denote the function
\[ f(a) = \begin{cases} 
1, & \text{if } a_k \leq a < a_{k+1}; \\
0, & \text{otherwise;} 
\end{cases} \]
and the convention that \( a_{K+1} = \infty \), I can write out the piecewise constant spline model as
\[ h(a) = \sum_{k=1}^{K} \gamma_k \mathbf{1}[a_k \leq a < a_{k+1}]. \]

By taking the piecewise polynomial corresponding to each knot as 0 before its knot and a linearly increasing function after, the model specializes to a piecewise linear spline model, a continuous function that has a constant derivative at all nonknots. By adding an additional basis function that is not associated with a knot, this piecewise linear spline model becomes a flexible approximation for any nonlinear function and is the main form I have used in representing age-specific hazards in the work to come. I can write out the piecewise constant specialization of the spline model as
\[ h(a) = \gamma_0 + \sum_{k=1}^{K} \gamma_k a \mathbf{1}[a \geq a_k]. \]

I find that in applications of this model it is useful to represent the piecewise linear spline in an alternative basis, where the model parameter \( \gamma_k \) represents the values of \( \log h(a_k) \) instead of the change in the slope at this point. This yields a more complicated set of basis functions, but it is not necessary to write out the basis functions explicitly.

Figure 3.2 shows the results of fitting spline models for age-specific hazards to simulated data to minimize the sum of the square differences between the predicted and observed values. When the piecewise constant model is fitted, it produces an age-specific hazard function consisting of a series of horizontal (constant) lines in each of the intervals between knots. Interval \( k \) has \( \gamma_k \) equal to the mean value of the simulated data between knots \( a_k \) and \( a_{k+1} \), which is quite a sensible choice.

A more favorable and flexible fit to the data is achieved by the piecewise linear spline model, which produces a continuous function of age as its prediction. In many cases, a piecewise linear fit of this type is sufficient to capture the nonlinearity in the data, and this will be the typical model.
for epidemiological rates in the second half of this book. It is possible to go further along this path of smoothing, however, and splines that have continuous derivatives and even continuous second derivatives are popular alternatives, achievable by simply choosing different piecewise polynomials for the basis functions.

![Figure 3.2](image)

**Figure 3.2.** Spline interpolation of simulated data. The true age-specific rate is piecewise log-linear, so none of the splines can represent it perfectly. The true age-specific rate and the models all have knot set \( \{0, 15, 60, 100\} \).

### 3.2 Choosing knots

To this point, I have taken as given the number and location of the knots in the spline model. However, selecting the number and location of the knots is an important task, and when working with sparse and noisy data, this can influence the model results substantially.

When data are abundant and age patterns are clear, models will not be very sensitive to the choice of knots, but when data are not abundant, or when the age patterns are not clear from the data, knot selection is a delicate part of the modeling process. In this setting, knot locations should be chosen a priori, based on the anticipated age groups available in the data as well as the anticipated changes in the epidemiology of disease.
being modeled as a function of age. For example, in a recent study looking at global trends in mean systolic blood pressure as a function of age, the modelers chose to use a cubic regression spline with knots located at ages 30 and 60. These choices reflect the expectation, based on literature and prior knowledge, that the behavior of mean systolic blood pressure as a function of age would be distinct in these intervals due to low blood pressure in young adults and to survivor effects in elderly populations. As another example, chapter 10 uses premenstrual syndrome to examine the effects of incorporating biological knowledge into the model as priors.

The approach of using expert knowledge to select the number and location of knots is a practical choice, but it is certainly not the only approach. Much literature is devoted to the choice of knot locations and the number of knots. An important direction for future work is to remove the reliance on expert knowledge to inform knot selection. This could proceed through model selection or model averaging of models with a variety of knot locations, through techniques developed in the adaptive regression spline literature, or by leaving spline models altogether and using Gaussian processes or some similar nonparametric model for the age pattern.

3.3 Penalized spline models

One approach to address the challenge of knot selection is to include many knots in the model and then also include a penalty function to discourage the model from using the additional knots when the data do not call for them. This penalized spline model can be formulated in a Bayesian framework by introducing a prior that represents the belief that, in the absence of evidence, the age pattern does not vary. Mathematically, I have formulated this as a penalty on the root mean square of the derivative of the age-specific rate $h(a)$:

$$\left[\int_{a=a_1}^{a_K} \|h'(a)\|^2 dw(a)\right]^{1/2} \sim N(0, \sigma^2).$$

This introduces an additional model parameter, $\sigma$, which can be viewed as a hyperprior and controls the amount of smoothing that the penalty creates.

For the piecewise linear penalized splines that will be used most frequently in the second half of this book, the derivative of $h$ is constant between knots, so, with equal weighting for smoothing at all ages, the integral
above simplifies to the following:

\[
\int_{a=a_1}^{a_K} \|h'(a)\|^2 \, da = \sum_{k=1}^{K-1} \left[ \frac{h(a_{k+1}) - h(a_k)}{a_{k+1} - a_k} \right]^2 (a_{k+1} - a_k) = \sum_{k=1}^{K-1} \frac{[h(a_{k+1}) - h(a_k)]^2}{(a_{k+1} - a_k)}.
\]

Figure 3.3 shows the effect of increasing the smoothing parameter \(\sigma\) when many more knots than necessary have been included in the model. Without smoothing, including many knots leads to estimates that are overly uncertain and wiggly. Smoothing, in the form of a quadratic penalty on the derivative of the age pattern, allows many knots to be included. But too much smoothing, for example, \(\sigma = 0.005\) in this case, results in a model that does not reflect true patterns in the data.
3.4 Augmenting the spline model

There are a few ways to augment the spline model that are useful when modeling age-specific rates. Since the epidemiological rates I have modeled are always nonnegative, I have parametrized the spline in terms of the log of the knot values, so that \( h(a) \) is a piecewise linear spline model with knots \( a_1, \ldots, a_K \), and

\[
h(a_k) = e^{\gamma_k}.
\]

To fit the model in a Bayesian framework, I have defaulted to giving these \( \gamma_i \)'s “weakly informative” priors,

\[
\gamma_i \sim \text{Normal}(0, 10^2).
\]

This has very little effect on the posterior distribution but makes the prior “proper” and also helps with algorithm convergence in some instances. Where relevant expert knowledge is available, I can replace this with a more informative prior (this idea is elaborated in chapter 4).

Finally, to deal with the order-of-magnitude differences of age-specific rates, I have applied the smoothing penalty to the logarithm of the rate rather than to the rate itself. This creates an additional complication, however, because the informative priors often say that rates are 0 for certain ages. To avoid a situation where the smoothing penalty includes the log of zero, I have rounded up any \( \gamma_i \) values that are below 10 times the mean rate. The approach is operationalized as a penalty term in the prior:

\[
\|h'\| = \sqrt{\sum_{k=1}^{K-1} \frac{[\max(\gamma_k, \gamma_{\text{min}}) - \max(\gamma_{k+1}, \gamma_{\text{min}})]^2}{(a_{k+1} - a_k)(a_K - a_1)}} \sim \text{Normal}(0, \sigma^2),
\]

where

\[
\gamma_{\text{min}} = \log \left[ \left( \sum_{i=0}^{K} e^{\gamma_i}/10 \right)/K \right].
\]
3.5 Summary and future work

Taken all together then, the model for an age-specific hazard function that will be used in this book is

\[ h(a) = \sum_{k=1}^{K-1} 1[a_k \leq a < a_{k+1}] \left( \frac{a - a_k}{a_{k+1} - a_k} e^{\gamma_k} + \frac{a_{k+1} - a}{a_{k+1} - a_k} e^{\gamma_{k+1}} \right), \]

\[ \gamma_k \sim \text{Normal} \left( 0, 10^2 \right), \]

\[ \|h'\| \sim \text{Normal}(0, \sigma^2). \]

The value of \( \sigma \) is a model parameter that will receive a very informative hyperprior: “slightly smooth” is represented by \( \sigma = 0.5 \), “moderately smooth” is represented by \( \sigma = 0.05 \), and “very smooth” is represented by \( \sigma = 0.005 \).

It is a limitation of this approach that the modeler must make a selection of the number of knots, location of knots, and level of smoothing, and developing computationally feasible alternatives which rely on the data alone is an important direction for future work.
Chapter 4

Expert priors on age patterns

Abraham D. Flaxman

When dealing with sparse and noisy data, it is sometimes necessary to include additional expert knowledge on the age pattern of epidemiological rates. For example, data sparsity can take the form of a lack of information about age-specific hazards of disease in children. In diseases that are rare or nonexistent in children, the fact that incidence is effectively 0 before a certain age is known by disease experts but not represented in the data collected by systematic review.

A benefit of the Bayesian methods that will be used to fit these models is the conceptual and practical simplicity of adding additional information to the age pattern model. This is implemented by choosing a more informative prior distribution. For example, if the epidemiology of a disease is such that the incidence level must be 0 before age $a_k$, this can be incorporated by replacing the weakly informative prior by the conditional probability density with this constraint included.

Three classes of additional information will come up frequently in the applications later in this book: level bound priors, level value priors, and monotonicity priors. This section describes how each can be implemented as an informative prior on the age pattern model.

4.1 Priors on level

Informative priors on the level of the age pattern seem simple at first but may have unintended effects. A prior on the level value for certain ages says precisely that the age pattern should have that value for those ages. For example, figure 4.1 shows the effects of adding a prior where the age-specific
hazard function takes values extremely close to 0.1, 0.5, or 1.0 from age 0 to 15.

![Figure 4.1](image)

**Figure 4.1.** An informative prior on the level of $h(a)$ for interval $0 \leq a < 15$ changes the estimated rate dramatically for $a$ between 20 and 60 and even leads to different estimates for $a = 100$.

These priors are implemented as “hard-soft constraints.” For a value $v$ on age range $(a_0, a_1)$, the value of the spline model is replaced with the level value for the age range (which I call a hard constraint), and the prior density on the spline is augmented with a penalty term for the offset log difference between the level value of the unconstrained spline and $v$ (which I call a soft constraint). The offset log difference penalty has the form

$$\log (h(a) + \epsilon) \sim \text{Normal} \left( \log (v + \epsilon), \sigma^2 \right),$$

where $h(a)$ is the age-specific hazard function, $\epsilon = 10^{-6}$ is the offset to avoid taking the log of 0, and $\sigma = 0.01$ is the magnitude of the penalty. In Bayesian terms, this encodes the belief that the spline is expected to be
within 1% of the expert level value, provided the level value is not too close to 0.

A similar sort of expert knowledge on the plausible bounds on level is also useful, both in modeling noisy data and in increasing the numerical stability of estimation algorithms. Again, however, the implications of such a prior can be unexpected. Figure 4.2 shows the effects of three different upper bounds on the spline estimation from the same data set as the previous figure.

![Figure 4.2](image-url)

**Figure 4.2.** An informative prior on the upper and lower bounds of the age-specific hazard function $h(a)$ changes the estimated hazard function dramatically for ages where the data are outside the bounds. For ages where the data are inside the bounds, the estimates are also affected, but to a lesser degree.

Like the level value prior, this prior is also implemented as a hard-soft constraint. If the level bounds are $\ell_0 \leq h(a) \leq \ell_1$, there is a hard constraint that replaces the spline with a clipped version, $h^c(a) = \min\{\max\{h(a), \ell_0\}, \ell_1\}$,
4.2 Priors on monotonicity

One common expert prior on age patterns is a strong belief that the function is increasing or decreasing over a certain age range. Mathematically speaking, these are priors on the sign of the derivative of the age pattern. For example, these priors can be implemented efficiently in Bayesian Markov chain Monte Carlo (MCMC) computation by conditioning on the differences of the age-specific hazard function \( h(a) \):

\[
h(a) \geq h(a + 1) \text{ for } a : a_s < a < a_e.
\]

The results of using such a prior are shown in figure 4.3. When the prior is contrary to the data, the estimate will be as close to the data as possible while respecting the prior. For example, the age-specific hazard function marked with triangles is the result of a prior belief that the age pattern increases from age 0 to 50 when confronted with data, shown as x-shaped marks that clearly decrease over this age range.

For computational efficiency, the increasing and decreasing constraints are implemented as soft constraints. For a constraint that the function is decreasing between \( a_s \) and \( a_e \), I include the penalty

\[
\min \{ \max [\log h(a + 1) - \log h(a), 0], 1 \} \sim \text{Normal}(0, \epsilon^2)
\]

for a small value of \( \epsilon \), like \( \epsilon = 10^{-6} \). This has a fully Bayesian interpretation, encoding a belief that a decreasing age pattern is expected and an increase of more than approximately 0.0003% is very surprising.

An area for future work comes from another common expert belief: that the age pattern is unimodal. This is conceptually clear, but computationally it has proven more difficult to realize than monotonicity. While the monotonicity constraint maintains log-concavity of the posterior distribution (if it was log-concave to start with), a straightforward implementation of a unimodality constraint will result in non-log-concave posterior distribution, even if everything else is well behaved. This suggests that the difficulty in fitting such models is inherent in the local step method of the MCMC algorithms I have been using. Perhaps an alternative approach such as the
4.3 PRIORS ARE NOT JUST FOR SPLINES

Figure 4.3. The expert belief that the age pattern is increasing or decreasing across an age range can also be implemented as a Bayesian prior.

population Monte Carlo algorithm would be more successful. Alternatively, there are some approximations of the unimodality constraint that may be easier to optimize over. (32)

4.3 Priors are not just for splines

Until now, all of the age patterns have been described as spline models, $h_a$. The compartmental model in section 7.2 will substantially expand the age pattern with many derived quantities such as the age-specific prevalence and the relative mortality risk. These quantities are derived in chapter 7. All three of the expert priors developed in this chapter are applicable to any age-specific function derived from the compartmental model. Most importantly, the age-specific prevalence can be augmented with expert priors on level values (e.g., birth prevalence is 0), level bounds (e.g., no population has
prevalence above 10%), and monotonicity constraints (e.g., prevalence in increasing as a function of age). Relative mortality risk is another derived quantity for which experts often have strong priors.

However, this sort of modeling requires care. The system dynamics model enforces a precise consistency between the different epidemiological rates, and making strong assumptions about one will have implications for others. Sometimes these implications are counterintuitive.

As a practical matter, I recommend that modeling begin with as few assumptions about the level and slope of the age pattern as possible which expert priors may be added one at a time. The benefit of this is threefold. First, fitting the model without all the available expert knowledge allows the data to speak. If the estimates confirm the expert belief, that is reassuring, and if they show the opposite, that is interesting. Second, the MCMC algorithm has a pitfall: nonconvergence. A quick way into this pit is introducing inconsistent expert priors, for example, decreasing prevalence and prevalence of 0 at age 0. By adding in expert priors one at a time, any inconsistencies that caused nonconvergence will be more easily identified. Third, as with any model that produces estimates from sparse and noisy data, it is essential to conduct a sensitivity analysis to understand how influential modeling assumptions are on the results. The gradual addition of expert priors will provide a starting point for this sensitivity analysis, showing which expert priors are essential to obtaining reasonable results and which are not as critical.

4.4 Empirical priors on age patterns

There is one additional type of level prior worthy of a separate exposition. It is the one that I have used to implement the empirical Bayes approach described in section 8.6. This computational convenience permits decomposing estimation of globally heterogeneous age patterns into subcomputations that can be run in parallel. It proceeds in two stages. Stage one estimates the mean and standard deviation for all regions assuming the age pattern is globally homogeneous. Stage two estimates a posterior distribution for each region independently, using the results from stage one as an empirical prior. This information is included in the age-specific hazard model through a penalty of the form

\[ h(a) \sim \text{Normal}(\mu_{\text{prior}}(a), \sigma^2_{\text{prior}}(a)) \].
Since the model must cope with the order-of-magnitude differences of age-specific rates, it can be more robust to use an empirical prior relating the offset log-transformed rates:

\[
\log (h(a) + \epsilon) \sim \text{Normal} \left( \log (\mu_{\text{prior}}(a) + \epsilon), \left( \frac{\sigma_{\text{prior}} + \epsilon}{\mu_{\text{prior}} + \epsilon} \right)^2 \right).
\]

4.5 Summary and future work

This chapter has introduced and demonstrated the effects of level value priors, level bound priors, and monotonicity priors, all of which can be applied to the spline models from chapter 3, as well as the age-specific rates that will be derived from the compartmental model in chapter 7. These priors can be very helpful when developing estimates for conditions with very sparse and very noisy data. However, these model assumptions must be justified, and future work is necessary in developing methods and procedures for conducting comprehensive sensitivity analyses and guaranteeing that the results are not unduly influenced by the choice of priors.

The application of empirical priors for age patterns is another area where future work will be important. There are alternative formulations for translating the estimates from a global model into priors for a regional model, such as a multivariate normal distribution with an empirically derived variance-covariance matrix. It remains to be seen if these more complex priors provide enough benefit to justify their use.

A prior related to the monotonicity priors which may be called a unimodality prior has often been requested during this work. Although it is straightforward to formulate the mathematical requirement that an age-specific hazard function have a single local maximum, all naive implementations have proven challenging for MCMC optimization. Developing a computationally tractable unimodality prior is an additional direction for future work.
Chapter 5

Statistical models for heterogeneous age groups

Abraham D. Flaxman

With a full development of statistical rate models for a single age group behind us, and the mathematical model for an age-specific rate function laid out as well, this chapter turns to a peculiar feature of population rate metaregression: the wide variety of age groups reported in the literature.

A typical example of the heterogeneity in age groups is shown for the systematic review results on atrial fibrillation (AF) prevalence figure 5.1. The midpoint of the age group is scattered against the width of the age group. Simply put, there is no standard set of age groups for AF research, and different studies report results with different age groups. Unfortunately, this phenomenon is far from unique to AF.

This variation in reporting would not be problematic if it were possible to access to the microdata from all the systematic review studies. For example, using microdata from a national health information system or from a demographic household survey, it would be possible to tally the prevalence rates by single-year age groups. Although each individual rate gathered in this way would have high variability, the rate model from chapter 2 combined with the spline model for an age-specific hazard function from chapter 3 would work together to produce an estimate that is as uncertain as it should be.

Although reanalysis from microdata is occasionally implemented in a GBD study, it is often not an option. I expect microdata reanalysis to become more frequent in national and subnational settings. In the more common situation where rate microdata are not available, the rates cannot
be recast into homogeneous age groups, and an alternative approach is needed.

This setting is in some ways similar to the settings where interval regression methods are applied in econometrics. However, I have more information about the structure of the data that I can leverage in my model.

I have considered several statistical approaches, and they will be compared and contrasted in this chapter. Before getting into the details, however, it is worthwhile to examine theoretically the way that age-grouping functions.

I begin with a simple mechanistic model of the age-grouping process. A study conducts some sort of measurement on a population of individuals who are all of different ages, and then the epidemiological rate or rates of interest are tallied for age groups selected in some context-dependent manner. If the study was a prevalence study using a full census sample, for example, and if I use \( r_{a_0,a_1} \) to denote the rate for age group \((a_0,a_1)\) and \( n_{a_0,a_1} \) to denote the subpopulation size of age group \((a_0,a_1)\), then the identity

\[
  n_{a_0,a_2} = n_{a_0,a_1} + n_{a_1,a_2}
\]

says nothing more complicated than that the size of the subpopulation of
age at least $a_0$ and less than $a_2$ is the sum of the size of the subpopulation between ages $a_0$ and $a_1$ and the size of the subpopulation between ages $a_1$ and $a_2$. Applying the same observation to the parts of these subpopulations that have the condition of interest yields the following identity:

$$r_{a_0,a_2} = r_{a_0,a_1} \frac{n_{a_0,a_1}}{n_{a_0,a_2}} + r_{a_1,a_2} \frac{n_{a_1,a_2}}{n_{a_0,a_2}}.$$ 

In a limiting case of a very large population with very fine age intervals, this becomes

$$r_{a_0,a_2} = \int_{a=a_0}^{a_2} r_{a,a+da} \frac{n_{a,a+da}}{n_{a_0,a_2}} da.$$ 

Undoubtedly, all real studies are more complicated than this full census of prevalence, but this is a starting point for conceptualizing where age-grouped rates come from. Roughly, they are integrals over instantaneous rates for infinitesimal age groups.

## 5.1 Overlapping age-group data

This section uses graphical statistics to explore an example of overlapping age-group data collected in systematic review. The primary way I like to display overlapping age-group data with horizontal lines on a plot of age versus rate value, as shown in figure 5.2. The level of the bars shows the rate value, while the width of the bars shows the range of ages included in the age group. It is often informative to augment these lines with error bars that show the uncertainty reported for each rate value, but for this section I have left out the representation of uncertainty to keep the plots as simple as possible.

Each of the horizontal lines in figure 5.2 can be represented as a triple $(a_s, a_e, r)$, where $a_s$ is the starting age of the age group, $a_e$ is the ending age of the age group, and $r$ is the rate observed for this age group.

A brief word about $a_e$ is in order here. Often in the epidemiological literature, the ending ages are described in a unit-dependent fashion, for example, age group 10–14. This is intended to mean from the first day of age 10 to the last day of age 14. However, this notation can be a hindrance when dealing with age resolution finer than 1 year, a situation that comes up when studying neonatal conditions. For this reason, I prefer the approach
Figure 5.2. The systematic review of the descriptive epidemiology of atrial fibrillation included 155 observations of disease prevalence for the USA. The prevalence level and age group of each observation are shown as a horizontal bar, with the position of the bar along the y-axis representing the prevalence level and the endpoints along the x-axis representing the start and end of the age group. The data show heterogeneity by age that is typical for these systematic review results and that clearly increases with age.

that takes the end age of the interval to be the first age when an individual is no longer part of the group. In the case above, I would say $a_e = 15$.

With a firm understanding of the sort of overlapping age-group data that arise in systematic review, I now turn to developing and analyzing a series of models for the meta-analysis of the data. I will consider five: the midpoint model, the disaggregation model, the midpoint-with-covariate model, the age-standardizing model, and the age-integrating model. The age-standardizing model is the balance of theoretical foundations, practical implementability, and empirical success that is used in the second half of this book.

5.2 Midpoint model

The simplest approach to modeling data with heterogeneous age intervals is to apply each rate measurement to the midpoint of the age group it
measures. This is trivial operationally, but it is also theoretically justified through a “trapezoidal rule” integration.

In practice, this approach is quite accurate for modeling a disease rate that changes slowly as a function of age. However, it becomes inaccurate when modeling rates that change more rapidly. The typical setting in applications in the second half of this book will include both a few studies that focus on age patterns and hence have narrow age groups and also many studies that focus on other aspects of disease epidemiology. Thus, when considering how these models are inaccurate, the relevant setting is where there are a few small age-group studies and many large age-group studies.

Mathematically, the formulation is as follows: let \( h(a) \) be a process model for the age-specific function (e.g., a spline model from chapter 3 or the age-specific prevalence function derived from the solution to the system of differential equations from section 7.2), and let \( p(r, n | \mu, \rho) \) be a data model for the observed level (e.g., the probability density function for the negative-binomial rate model from chapter 2). Then the likelihood of an observation of rate \( r_i \) with effective sample size \( n_i \) for age group \((a_{si}, a_{ei})\) is simply \( p(r_i, n_i | h(a_{si} + a_{ei}/2), \rho) \). Equivalently, in “blackboard notation,” using \( D(\mu, \rho; n_i) \) to denote the rate model distribution, I can write

\[
    r_i \sim D(h(a_i), \rho; n_i), \\
    a_i = \frac{a_{si} + a_{ei}}{2}.
\]

This formulation will be convenient for comparison with the other models of age groups to come.

To understand how accurately age-group models like the midpoint model can estimate, I used simulation. The precise details are deferred until section 5.6, but since this simulation is also used for the figures that follow, I will describe it briefly here. First, I selected an age-specific hazard function as ground truth. Then I generated noisy measurements from a mixture of regularly spaced 10-year age groups and uniformly random age groups. For each measurement, I chose a random population structure and integrated the true age-specific hazard to find the true rate for the age group. Then I sampled from a negative-binomial distribution with this true rate as the mean and a fixed overdispersion parameter to obtain noisy data, which I used in the age-group model. Since ground truth is known in this simulation, I can compare the model estimates to the truth graphically as well as quantitatively.
CHAPTER 5. STATISTICAL MODELS FOR HETEROGENEOUS AGE GROUPS

Figure 5.3 compares the estimate produced by the midpoint model to ground truth through simulation using two different age-specific hazard functions as ground truth. When the age-specific hazard varies little as a function of age, as shown in panel (a), the estimated hazard function is quite accurate. But when the age-specific hazard function varies substantially, as shown in panel (b), the estimate is biased.

Figure 5.3. The midpoint model, a conceptually simple approach to dealing with data with heterogeneous age groups, simply attributes the observation to the midpoint of the age group. Panel (a) shows the model applied to an age-specific hazard function that does not vary a great deal across ages; the midpoint model is an accurate fit. Panel (b) shows the model applied to a more variable age-specific hazard function; the midpoint model overcompresses the estimates.

5.3 Disaggregation model

An alternative to the midpoint model that seems appealing but has some downsides is what I call disaggregation. To understand the disaggregation approach, imagine the simple reanalysis that I could do if microdata were available (as described at the beginning of this chapter). If I had access to
the individual measurements that went into the calculation of the disease rate found in systematic review, I could do a reanalysis with any age grouping I wished. I could calculate rates for single-year age groups and be sure that the age pattern does not change substantially during the grouping.

The microdata from rates found in systematic review are rarely available, however. The disaggregation approach is a simple attempt to impute what the rates for the desired age grouping would be if the microdata were available. This requires taking into account the increased variation that would be found if a study of the same size was reported for finer age groups.

Without any additional information, rate data reporting a level of \( r \) for a population with effective sample size \( n \) for age group \((a_s, a_e)\), that is,

\[
X = (r, n, a_s, a_e)
\]

can be disaggregated into \( A = a_e - a_s \) rows of data, \( X_1, X_2, \ldots, X_A \), with

\[
X_a = \left(r, \frac{n}{a_e - a_s}, a, a + 1\right), \text{ for } a = 1, 2, \ldots, A.
\]

Disaggregation can be interpreted as a data-preprocessing step, and these disaggregated data can be fed into the midpoint model from the previous section to produce a comprehensive estimate of the rate as a function of age. However, this model has some unintended negative features when large age intervals are disaggregated. Because it ignores the correlation of disease levels with age, it tends to overcompress age patterns at young and old ages, as shown with simulated data in figure 5.4.

### 5.4 Midpoint model with group width covariate

An alternative method, which I consider more “statistical” in its approach, is to add the width of the age group as a covariate into the midpoint model. This model takes the form

\[
r_i \sim D(\mu_i, \rho; n_i),
\]

\[
\mu_i = h\left(\frac{a_{si} + a_{ei}}{2}\right) + \theta(a_e - a_s).
\]
CHAPTER 5. STATISTICAL MODELS FOR HETEROGENEOUS AGE GROUPS

Figure 5.4. The disaggregation approach to two simulated data sets. In (a) the age groups are sufficiently fine-grained and homogeneous, and disaggregation is a successful approach. But in (b) with even slight heterogeneity, the model estimates are overcompressed.

This addresses the shortcomings of the disaggregation approach indirectly, and the indirect nature has positives and negatives. This method does not explicitly connect the large age interval to the small age interval but instead allows the data to inform the relationship. On the other hand, it posits that the data-driven relationship between the rates for studies with the same midpoint but different age groups is a linear relationship. In contrast, the mathematical model developed at the beginning of this chapter is nonlinear in a specific and mechanistically known way. Figure 5.5 shows the results of applying the midpoint-covariate model to simulated data.
5.5. **AGE-STANDARDIZING AND AGE-INTEGRATING MODELS**

An even more complicated approach, both conceptually and computationally, is to average across the age interval explicitly in the statistical model:

\[ r_i \sim D(\mu_i, \rho; n_i), \]
\[ \mu_i = \int_{a=\alpha_i}^{a=\beta_i} h(a)dw_i(a), \]

where the integration \(dw_i\) is weighted according to population structure.

This has the theoretical appeal of matching the generative model above but the drawback of being slower computationally and less stable numerically. It also has a major piece left unspecified: the selection of the age weights for the integration. There are two sensible approaches to age-weight selection, which I call the *age-standardizing model* and the *age-integrating model*. The age-standardizing model uses a common age pattern \(dw_i(a) = dw(a)\) for all studies, while the age-integrating model uses

---

**Figure 5.5.** The midpoint-covariate model applied to two simulated data sets, where ground truth is known. Although this approach is appealing theoretically, the added flexibility of the covariate model does not add much value in the simulation study.
the best estimate available of the age pattern of the study population in each observation. The age-standardizing model is faster, due to a computational optimization only possible when the $d_{w_i}$ are the same for all $i$, but the age-integrating model is appealing on theoretical grounds, because it can use more information. However, it is not certain that using this information will make the end results any more accurate, because the age pattern of the study population is not always known with great certainty, and then it is necessary to assume that it matches the national age pattern for the country-years where the study was conducted. In the case of remission and mortality studies it is even more complicated to estimate the study population age pattern, since it is not the same as the national population age pattern but modulated by the age pattern of disease prevalence. Figure 5.6 shows the results of the age-standardizing model on simulated data.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5_6.png}
\caption{The age-standardizing model applied to simulated data with a known age-specific rate function as ground truth. The results in panel (a) show that the model recovers the true age pattern quite precisely. Panel (b) shows that the results are still accurate when the data generation procedure is even more noisy.}
\end{figure}
5.6 Model comparison

Figure 5.7. A comparison of 4 models for heterogeneous age groups shows that the age-standardizing model comes closest to recovering the truth. This corresponds to the results of the simulation study presented in table 5.1.

This section provides a comparison of the approaches to age-group modeling. An appropriate comparison of these approaches is somewhat difficult to develop. One approach is through simulation, where a data set is simulated from known ground truth (figure 5.7). This allows the estimates to be compared to “true” values, but this risks inappropriate model selection due to inaccurately choosing the distribution of the simulated data. Another approach is cross-validation, where data from systematic review are split into mutually exclusive training and test sets, and the model is fitted to the training set and used to predict the values in the test set. Naively holding
out 25% of the data does not address the exact topic of interest, however, since it determines which model predicts rates of all age groups, and I am really only interested in predicting the age groups with small widths accurately. It would be preferable to hold out only data with small-width age groups from large representative subpopulations. Unfortunately, there are rarely enough data to do this, especially in all the settings that come up in disease modeling.

I have taken a pragmatic approach, evaluating by means of the natural simulation described below. Future work, based on more sophisticated simulation scenarios or based on carefully designed holdout cross-validation, is necessary to further understand the trade-offs between these alternative methods.

The data simulation procedure I used is the following:

- Choose age intervals for 30 rows of data; for \( i = 1, \ldots, 10 \), \((a_{si}, a_{ei}) = (10(i - 1), 10i)\), and for the remaining 20 intervals, choose the age interval width uniformly at random from \([1, 100]\) and choose the midpoint of the age interval uniformly at random from ages compatible this age range.

- Choose the effective sample size \( n_i \) for each row uniformly at random from \([10^2, 10^4]\).

- Choose an age-specific population structure for each row of data, with the form \( w_i(a) = e^{\beta_i a} \), where \( \beta_i \) is drawn from a normal distribution with mean 0 and standard deviation \( \frac{1}{10} \).

- Calculate the true rate value for each age interval,

\[
 r_i^{\text{true}} = \sum_{a=a_{si}}^{a_{ei}} \mu_{\text{true}}(a)w_i(a),
\]

where

\[
 \mu_{\text{true}}(a) = \exp \left\{ \frac{3(a - 35)^2}{1000} + \frac{a - 35}{100} \right\}.
\]

- Choose an observed rate value, based on a negative binomial distribution: \( r_i n_i \sim \text{NegativeBinomial}(r_i^{\text{true}}, \delta_{\text{true}}) \), where \( \delta_{\text{true}} = 5 \).

Table 5.1 shows the median results of fitting this simulated data for a variety of models. The age-standardizing models is superior in all metrics of fit quality.
### 5.7 Summary and future work

Table 5.1. Median results for 100 replicates of the simulation study comparing age-specific rate estimates from 5 models of age-grouped data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Bias (pp)</th>
<th>MAE (pp)</th>
<th>PC (%)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midpoint</td>
<td>3</td>
<td>6</td>
<td>60</td>
<td>32</td>
</tr>
<tr>
<td>Disaggregation</td>
<td>-1</td>
<td>22</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>Midpoint-covariate</td>
<td>4</td>
<td>6</td>
<td>72</td>
<td>37</td>
</tr>
<tr>
<td>Age-standardizing</td>
<td>0</td>
<td>3</td>
<td>75</td>
<td>34</td>
</tr>
<tr>
<td>Age-integrating</td>
<td>0</td>
<td>3</td>
<td>70</td>
<td>35</td>
</tr>
</tbody>
</table>

*Note:* Bias is the mean of truth minus predicted measured in percentage points (pp), median absolute error (MAE) is the median of absolute difference between truth and predicted, probability of coverage (PC) is the fraction of truth falling within 95% uncertainty interval of prediction, and time is the computation time.

This chapter has developed and compared several alternative approaches to dealing with heterogeneous age groups in systematic review data. Although the age-standardizing model has been preferred on theoretical grounds and simulation study, future work comparing the performance of these approaches at out-of-sample prediction is necessary to confirm this choice. Additional approaches to age group heterogeneity based on alternative statistical or machine learning methods may also be worthwhile.
Chapter 6

Covariate modeling

Abraham D. Flaxman

As frequently mentioned in previous sections, the epidemiological data on disease morbidity collected in systematic review is often very sparse and very noisy. Covariate modeling is a method to explain the variation in noisy data in terms of demographic, epidemiological, and study-specific variables. This is often challenging because there is no particularly explanatory variable available and also because the data are very sparse.

In my metaregression model of disease in populations, covariate modeling has two distinct goals. One is to explain the bias and variation of the noisy measurements of epidemiological rates. For example, covariates can be used as a mechanism for data-driven “cross-walks” to convert between alternative diagnostic methods that have different sensitivities, and covariates can also be used to objectively down-weight data that come from a noisier source such as nonrepresentative subpopulations when they are not systematically biased above or below the mean.

The other goal in covariate modeling is to increase the accuracy of out-of-sample predictions. This is accomplished by modeling the relationships between the disease parameters of interest and the explanatory covariates. The modeled relationships are then used to extrapolate predictions for the disease parameters to geographic regions where covariate data are available but where no or few direct measurements have been made.

In covariate modeling, there is often a distinction made between “fixed effects” and “random effects.” Bayesian approaches, such as hierarchical modeling, blur this distinction. To make the nomenclature more complicated, different methodological traditions of covariate modeling have opposite concepts of what is fixed and what is random in effects.
CHAPTER 6. COVARIATE MODELING

For this metaregression framework, I have used fixed and random effects in different ways, which makes them easy to keep separated. While I have used fixed effects for covariates that vary by study or by country-year, I have used random effects to model only indicator covariates for geographic units.

Sometimes I have constrained the random effects to sum to 0 at each level of a geographic hierarchy, which is an extension of the traditional meaning of random effects in linear regression, where the population mean of a random effect is 0. In other models, it is sufficient to use a prior with mean 0 for independently for each random effect, as is the common approach in Bayesian modeling. In either case, my random effects always have a hyperparameter for the dispersion, which allows the model to infer how dispersed the random effects are between geographic regions and hence to quantify the uncertainty in the geographic regions for which no data are available.

The fixed effects thus model variation between measurements that can be explained, while the random effects model true variation between measurements for which we have no explanatory covariates. Distinct from both of these is the model for sampling and nonsampling variation in the measurements, which is implicit in the rate models developed in chapter 2. In the case of a negative binomial rate model with random effects, for example, the model must distinguish between true variation from country to country and nonsampling variation. This is quite a challenge when limited data are available.

I will develop all these concepts in the following sections of this chapter.

6.1 Cross-walk fixed effects to explain bias

A prototypical example comes from myocardial infarction (MI) incidence, where a variety of diagnostic tests are available. Different studies of MI incidence use different diagnostic criteria for case ascertainment. The newer class of tests, which are based on measuring levels of the blood protein troponin, are more sensitive than earlier methods, and this leads to variation in data with a clear explanation. Figure 6.1 shows simulated data with a covariate that has an effect like a troponin-based test might, raising the number of observed cases by 30%. By including an indicator variable as a covariate in each row of data, \( x_i = 1 \) if row \( i \) comes from a study that
used a troponin test, and \( x_i = 0 \) otherwise, I can fit a model that includes a parameter to cross-walk between studies using these two different case ascertainment criteria.

![Figure 6.1](https://example.com/figure61.png)

\textbf{Figure 6.1.} Simulated data set where different measurement techniques yield systematically different values. The data with \( x_i = 1 \) are on average 30\% higher than data with \( x_i = 0 \), and the covariate model recovers this difference accurately, with sufficient data.

This same approach can be applied to data on mental disorders gathered with different recall periods, an application that arises frequently in the meta-analysis of psychological disorders. For example, in measuring the population prevalence of bipolar disorder, many studies ask about symptoms in the past month, while many others ask about the past year. Figure 6.2 shows the data collected in systematic review for bipolar disorder, where past-year prevalence is higher than past-month prevalence because of the episodic nature of the condition.

In general, let the data collected in systematic review be denoted by tuples \((a_i, n_i, r_i, X_i)\), where \( a_i \) is the age group, \( n_i \) is the effective sample size, \( r_i \) is the observed rate value, and \( X_i \) is a vector of covariate values. Then, using \( D(\pi, \rho; n_i) \) to denote the rate model, the fixed-effects covariate
The parameter $\beta$ represents the effect coefficients for the fixed effects, and because the data are often sparse and noisy, it can help the stability of the computational algorithms to put a weakly informative prior on $\beta$, such as

$$\beta_j \sim \text{Normal}(0, 1^2) \text{ for } j = 1, \ldots, J.$$ 

Of course, if experts have beliefs about the sign or magnitude of the effect coefficient, this can be included as a more informative prior.

Two subtle choices are worth additional investigation in fixed-effects modeling: normalization and reference values. Both of these choices are known to influence the performance of computational algorithms. For example, nonnormalized covariates can produce nonconvergence in hill-climbing algorithms that work fine with normalized covariates. But because of the Bayesian priors and especially because of the consistency from the compartmental model, the choices are particularly important in this setting.

The term reference value is borrowed from fixed-effects modeling of categorical variables, where so-called dummy variables (0/1 indicators) are
6.2. Predictive fixed effects to improve out-of-sample estimation

introduced for all but one category. When all the dummy covariates are set to 0, the model produces predictions for the reference category. In the formulation above, the analogous notion occurs when \( X_i = (0, 0, \ldots, 0) \). Then the expression for \( \mu_i \) simplifies to

\[
\mu_i = h(a_i) e^{\beta_0} = h(a_i).
\]

It is this \( h \) that is used as the age-specific rate function in the compartmental model (as developed in chapter 7), so the consistency between incidence, prevalence, remission, and mortality is enforced at the reference values.

Because the reference values are consistent, they must be chosen with care. For example, in the case of MI above, where some studies used troponin-based diagnostics and some did not, the reference value should be with troponin tests, because this is considered to be more accurate.

A concrete example using the bipolar disorder data can make this clearer. Chapter 14 provides another example, while chapter 19 develops the consistent model for bipolar disorder in detail, which is used here in two variations: when the past-year prevalence is used as the reference value, and when the past-month prevalence is used as the reference value. This changes the predicted prevalence, of course, but it also changes the predicted incidence (for which little data are available). Figure 6.3 shows how the alternative reference values change the incidence estimate in this case.

Normalization is also important, although it does not affect consistency. It is important for stability of numerical algorithms and also because the prior on the effect coefficient must be matched to the scale of the covariate. Normalizing continuous covariates to have variance 1, for example, means that the prior of \( \beta \sim \text{Normal}(0, 1^2) \) is weakly informative. If a continuous covariate had variance 0.0001, the same prior on \( \beta \) would be very informative.

6.2 Predictive fixed effects to improve out-of-sample estimation

In addition to study-level covariates, like the cross-walks in the previous section, covariate modeling can be used at the country level to use relationships measured in-sample to improve estimation of true regional variation out-of-sample. Mathematically, the setting is identical, where a country-level covariate matrix \( X'_i \) holds the value of the covariates, and an effect
Figure 6.3. The reference value for the past-year/past-month prevalence covariate has a substantial effect on incidence estimates. Because consistency is enforced at the reference value level, choosing reference values is an important modeling decision.

Conceptually, this deserves separate treatment, however, because the use and the results of country-level covariate modeling are quite different.

The benefit of using fixed effects to predict out-of-sample is clear when modeling an often fatal condition, like decompensated cirrhosis. Incidence of this disease is available from registries for some regions, but population-level mortality caused by the condition has been modeled in detail for all countries. By using the log of the age-standardized mortality rate as a covariate in the incidence model, it is possible to borrow strength from the mortality estimates to inform the incidence estimates. Chapter 15 explores this specific example in more detail.

This approach can also be helpful for covariates that are not as direct, for example, using gross domestic product as an explanatory covariate for estimating the prevalence of eating disorders, using estimates of age-standardized hepatitis C virus prevalence as an explanatory covariate for estimating prevalence of cirrhosis, or using an indicator for violent conflict as an explanatory covariate for estimating the prevalence of depression and anxiety disorder.

However, to use country-level covariates in this way requires having
a time series with country level data for each country and year of study included in the analysis. This data may be sparse and noisy itself, and often requires imputation for missing country or year values.

6.3 Fixed effects to explain variance

Fixed-effects modeling in the previous sections has focused on improving predictions of the mean of observed data. It is also possible to use fixed-effects modeling to explain the different levels of variation in different sources of data, which is the topic of this section.

To introduce this idea by way of example, consider the results of a systematic review for hepatitis C virus seroprevalence. This literature search excluded studies in subpopulations known to have systematic bias, such as studies of prevalence in intravenous drug users or paid blood donors. But it did collect measurements from studies in subpopulations that were not known to be systematically biased, for example, studies that used voluntary blood donors as the sample frame. This is clearly not the whole population, but as it is not known to be a biased sample, I would like to include it if possible. This is where using a fixed effect to explain variation is appropriate. The systematic review assigned a bias indicator $Z_i = 1$ to observations corresponding to the voluntary blood donors, as well as to observations from other studies of nonrepresentative subpopulations, such as mothers visiting antenatal clinics. Observations from studies of the general population received bias indicator $Z_i = 0$. Then I was able to introduce a fixed-effect coefficient analogous to that discussed in the previous sections, but modifying the overdispersion term of the rate model instead of the mean.

This procedure resulted in the following formulation:

\[ r_i \sim D(\mu_i, \delta_i; n_i) , \]
\[ \mu_i = h(a_i)e^{\beta X_i} , \]
\[ \delta_i = e^{\eta + \zeta Z_i} . \]

6.4 Random effects for spatial variation

Another important use of covariates is in handling nonsampling variation that cannot be explained. As I have mentioned repeatedly, the descriptive
epidemiological data available are often very noisy. Usually only a small part of this “noise” can be explained with covariates like those from the preceding section. And while the additional variation has no simple explanation in terms of differing diagnostic criteria or the like, there is structure in the variation. Countries in the North Africa and Middle East region have rates more similar to each other than to countries in the high-income North America region. And the high-income North America region as a whole is more similar to the Western Europe region than to the South Asia region. Capturing this spatial similarity when it exists is the goal in my random-effects modeling.

I will develop this approach to random-effects modeling by beginning with something very similar to the fixed-effects model. The random effects come, in part, through the use of additional priors, either modeling the dispersion of the effects as a parameter itself to be fitted from the data or going further to model the joint distribution of spatially neighboring effects to have sums equal to 0. For notation, let $U_i$ be a vector of random-effects covariates. This $U_i$ is a design matrix analogous to the fixed-effect covariate vector $X_i$ above, but with 0/1 values corresponding to the place in the spatial hierarchy to which observation $i$ refers.

In the GBD Study 2010, the spatial hierarchy is countries nested in regions nested in superregions, but in national or subnational analyses, the hierarchy will be different. This can be generically formulated using graph theory, where a directed tree (also known as an out-arborescence) encodes the hierarchical relationship structure with a root node connected by out-arcs to children on the first level of the hierarchy, which are each in turn connected by out-arcs to children on the next level of the hierarchy, and so on. A node is called the parent of any node it points to in this tree, and two nodes are called siblings if they share the same parent.

Analogously to the fixed-effects model above, the random effects apply a multiplicative shift to the age-specific rate function:

$$r_i \sim \mathcal{D} (\mu_i, \delta_i; n_i),$$

$$\mu_i = h(a_i)e^{\alpha U_i}.$$

The first difference between the fixed effects and random effects is in the priors on the effect coefficients. Instead of a weakly informative prior as above, the prior on $\alpha$ is itself part of the model, parametrized as

$$\alpha_j \sim \text{Normal} \left(0, \sigma^2_{\ell(j)}\right),$$
where \( \ell(j) \) is the level in the hierarchy of node \( j \), and \( \sigma_\ell \) is also a model parameter. To fit this model with Bayesian methods, we also need a prior on \( \sigma_\ell \) (a hyperprior), and because of the sparse and noisy nature of the available data, this often has to be somewhat informative. The truncated normal distribution

\[
\sigma_\ell \sim \text{Normal}_{0.05,5}(0.05, 0.03^2)
\]

is often an appropriate choice. It says that between-area variation of less than 5% is impossible and more than 15% is rare.

A second difference between the fixed effects and random effects that can help with MCMC convergence is the following modification to the joint prior distribution of \( (\alpha_j) \): for every node in the spatial hierarchy, I can constrain the random effects for all children of that node to sum to 0. Using notation from graph theory and letting graph \( H \) represent the hierarchy, this constraint can be formalized mathematically as

\[
\sum_{c \in N^+(p)} \alpha_c = 0, \text{ for all } p \in V(H)
\]

The zero-sum prior has important implications in consistent models, because as described above (and shown in figure 6.3), consistency is enforced at the reference level, which for random effects is \( U_i = 0 \). The zero-sum constraint has the benefit of reducing the number of dimensions in the parameter space, which is why it helps with MCMC convergence.

\section*{6.5 Covariates and consistency}

One of the most challenging theoretical issues in covariate modeling for integrative systems modeling is the interplay between the predictive covariates and the intercompartmental consistency. A simple example of the problem arises in a model of congenital abnormalities, where there is where there is birth prevalence, prevalence at older ages and mortality risk but not incidence and remission is zero. If covariates are used to shift predictions for the level of \( h_p \cdot h_f \) as well as the level of \( h_p \) and the level of \( h_f \), then consistency would require that \( \beta_{hpf}^i = \beta_{hp}^i + \beta_{hf}^i \).

This complication becomes even more pronounced in a model with nonzero incidence and remission. In the general case, it is not even clear that nonzero covariate effects exist that respect consistency.
To circumvent this challenge, I have used a multistage approach to fitting the model (see section 8.6), and at each stage of the process, there is a specific level of the hierarchical model where I have enforced the consistency conditions of the system dynamics model. All predictions from this stage apply only to this node and nodes lower in the hierarchy, and for the lower nodes, the predictions are not consistent. However, they are expected to be close to consistent, a hypothesis that must be investigated empirically on a case-by-case basis.

How does this work? Recall the covariate model formulation for predicting the rate for a given geographic area, sex, and year \((g, s, y)\):

\[
\pi_{g,s,y}(a) = h(a)e^{\alpha U_{g,s,y} + \beta X_{g,s,y}}.
\]

For the highest node of the hierarchy (also called the reference node and corresponding to geographic area, sex, and year \((g_r, s_r, y_r)\)), I simply apply a linear shift to each covariate in \(X\) and \(U\) so that \(X_{g_r,s_r,y_r} = 0\) and \(U_{g_r,s_r,y_r} = 0\). This simplifies to

\[
\pi_{g_r,s_r,y_r}(a) = h(a),
\]

and for any system of differential equations for which \(\{h_t(a), t = [T]\}\) are solutions, the predicted values for the age, sex, and year at the root of the hierarchy are also solutions.

An important direction for future work is to go beyond the multistage approach. This will probably require innovation in algorithms, because fitting multiple consistent models simultaneously is currently impractical.

### 6.6 Summary and future work

This chapter described the multiple ways covariates have been used in descriptive epidemiological metaregression: to explain bias, to improve the accuracy of out-of-sample prediction, to explain variance, and to measure unexplained variation. These different applications are all similar mathematically, but there is much subtlety in how each influences the model estimates.

In future work, it will be important to develop covariates that themselves include uncertainty, since many predictive covariates are themselves estimates. Similarly, methods that allow covariates with missing values will
be useful in future modeling efforts. The covariate modeling developed in the Cause of Death portion of the Global Burden of Disease Study 2010 benefited greatly from ensemble modeling methods,\(^{36}\) an additional approach that could be tried here as well.
Chapter 7

Prevalence estimates from other data types

Abraham D. Flaxman

To inform age-specific estimates of prevalence with data on other epidemiological parameters (such as incidence, remission, and mortality), this chapter introduces the framework that I call *integrative systems modeling* (ISM). ISM combines a mechanistic model of process with a statistical model of data. The foundations of ISM are best articulated in terms of *system dynamics modeling*, a discipline that originated in the fields of operations research and industrial engineering. This type of compartmental modeling is similar to infectious disease modeling and pharmacokinetic/pharmacodynamic (PK/PD) modeling. System dynamics modeling is a method to model the behavior of complex systems in terms of stocks, flows, and feedback loops. In short, *stock variables* quantify the amount of some material, mass, or population in a compartment at a particular moment in time, while *flow variables* quantify the rate of material moving into, out of, or between compartments. The scope of applications for system dynamics is enormous, and once you start thinking of systems in this way, it may seem that everything can be modeled as stocks and flows. This method has been applied to the study of complex systems in economics, politics, environmental science, and a diverse array of other fields.

Traditionally, there is a delineation between system dynamics modeling and statistical modeling: system dynamics aims to develop a *model of process*, while statistical approaches focus on developing a *model of data*. Models of process attempt to explicitly represent the mechanisms behind...
some system behavior (deterministically or stochastically), while models of data often explicitly avoid requiring such a theory. The advantage of using the system dynamics approach is that it can incorporate structural assumptions about the system. However, in many applications, the system dynamics model of process is not connected to data at all. On the other hand, in many statistical approaches, the domain-specific dynamics of the system under study are not incorporated in the model explicitly, and this exclusion may be intentional, to allow the data to speak for themselves. However, in the case of sparse and noisy data, data models could benefit from additional structure. ISM is a framework to bring together a model of process and a model of data for mutual benefit.

7.1 A motivating example: population dynamics

An example will help to make these concepts more concrete. We begin with the simplest of compartmental models, a single compartment with inflow, outflow, and no feedback, shown schematically in figure 7.1.

![Figure 7.1. A single-compartment model with inflow b and outflow m is one of the simplest examples of a compartmental model. Despite its simplicity, it is a useful model of population dynamics. In this application, b represents births, m represents mortality, while S represents the “stock” of population.](image)

Schematic diagrams of stock-and-flow models such as figure 7.1 are useful in understanding and communicating the structure of a model of process, but they are incomplete (at least, as commonly used in epidemiology; conventions in PK/PD differ). The full description is represented in the form of a system of difference equations or differential equations that specify precisely the relationship between the stocks and flows. The following differential equation fully specifies the one-compartment model:
7.2 SYSTEM DYNAMICS MODEL OF DISEASE IN A POPULATION

\[ \frac{dS}{dt} = b - m, \]
\[ b = h_b S, \]
\[ m = h_m S. \]

In this equation, the stock \( S \) changes continuously, increasing with birth hazard \( h_b \) and decreasing with mortality hazard \( h_m \). When \( h_b \) and \( h_m \) are constants with respect to time, this differential equation has a closed-form solution: \( S = S_0 e^{(h_b - h_m)t} \). When \( h_b \) and \( h_m \) are not constant with respect to time, the model does not necessarily have a closed-form solution, and many more time trends are possible for \( S \). Panel (a) of figure 7.2 shows the time trend of \( S \) when \( h_b \) and \( h_m \) are constant; panel (b) shows the time trend when they are changing.

The next section focuses on the specific application that is crucial to model-based metaregression: the system dynamics of a disease moving through a population, with flows that vary as a function of age.

7.2 System dynamics model of disease in a population

The key to combining data of different types is a two-compartment system dynamics model of process. The compartments contain the population susceptible to the disease (stock \( S \), for “susceptible”) and the population with the condition (stock \( C \), for “condition”). The population moves between these compartments following the arrows shown in figure 7.3, transitioning from \( S \) to \( C \) with incidence hazard \( h_i \), and from \( C \) back to \( S \) with remission hazard \( h_r \). The susceptible population flows out of the system with background mortality hazard \( h_m \), and the with-condition population flows out of the system with with-condition mortality hazard \( h_{m\text{with}} = h_m + h_f \). Here \( h_f \) represents the “excess mortality hazard” for individuals with the condition over individuals without the condition.

This “model of process” looks deceptively simple, and compared to the complex systems often developed in infectious disease modeling it is quite simple. However, it more flexible than it appears at first glance. All the parameters of the model are functions of age and time, which permits enough
flexibility to match well the myriad of available data about the descriptive epidemiology of disease. I will connect this model of process to data using the age-standardizing mixed-effects spline models from the previous chapters, and together they will provide the ISM for metaregression that incorporates data on parameters other than prevalence into a prevalence estimate.

As mentioned in the previous section, a schematic depiction of a compartmental model, such as figure 7.3, is not a complete description of the
7.2. SYSTEM DYNAMICS MODEL OF DISEASE IN A POPULATION

![Diagram of compartmental model](image)

Figure 7.3. The two-compartment model of process for disease in a population, which allows prevalence estimates to be informed by other types of data. Compartment $S$ contains the population susceptible to the disease, and compartment $C$ contains the population with the condition. Individuals move from $S$ to $C$ with incidence hazard $h_i$, and from $C$ to $S$ with remission hazard $h_r$. The susceptible population flows out of the system with without-condition mortality hazard $h_m$, and the with-condition population flows out of the system with with-condition mortality hazard $h_m + h_f$, where $h_f$ is the excess mortality hazard and represents quantitatively the increase in mortality for individuals with the condition.

System dynamics. To specify them completely requires a system of differential equations that correspond to the stocks and flows above and that more precisely represent their relationship as a function of time and age:

\[
\frac{d}{d \tau} S(a + \tau, t + \tau) = -(h_i + h_m)S + h_rC,
\]

\[
\frac{d}{d \tau} C(a + \tau, t + \tau) = h_iS - (h_r + h_m + h_f)C.
\]

The notation $S, C, h_i, h_r, h_m$, and $h_f$ denote the following functions of age
CHAPTER 7. PREVALENCE ESTIMATES FROM OTHER DATA

and time:

\[ S = S(a, t) = \text{stock of “susceptibles,”} \]
\[ C = C(a, t) = \text{stock of “with condition,”} \]

\[ h_i = h_i(a, t) = \text{incidence hazard for susceptibles,} \]
\[ h_r = h_r(a, t) = \text{remission hazard for individuals with the condition,} \]
\[ h_m = h_m(a, t) = \text{without-condition mortality hazard for susceptibles,} \]
\[ h_f = h_f(a, t) = \text{excess mortality hazard for individuals with the condition.} \]

In general, all these quantities are functions of both age \( a \) and time \( t \).

At this point in developing the model of process, it is valuable to pause and consider what descriptive epidemiological data may be collected in a systematic review of the literature and other sources, and how they relate to the stocks and flows, by using a suitable model of data.

For certain especially infectious diseases, for example, tuberculosis, cases identified in the health system are regularly reported to WHO. This yields data on disease incidence rates (which are not age-specific rates but crude incidence rates over all ages). The number of cases per year divided by the midyear population provides an approximate measurement of the incidence hazard \( h_i \) in figure 7.3.

Often it is prevalence that is directly measured, for example, through a household or telephone survey. This sort of research provides measurements of the form “\( k \) out of \( n \) respondents tested positive for the condition” (or said that they had the condition). Simple division or some more complicated survey analysis method then yields a measurement of the ratio of compartments in the stock-and-flow model above: that is, prevalence is equal to \( C/(S + C) \). Often this information will be stratified by age, sex, or both, and for many important diseases, the prevalence level will vary by orders of magnitude over the age range of the population.

Another sort of study that is sometimes available measures the relative mortality ratio of a disease, meaning the mortality rate in people with the disease and without the disease, and reports the ratio (sometime called the relative risk). This corresponds to a ratio of hazards in the model above, \( (h_m + h_f)/h_m \). Unfortunately, \( h_f \) is rarely measured directly, but the with-condition mortality hazard is sometimes reported instead of relative mortality risk, which can be represented as \( h_m^{\text{with}} = h_m + h_f \).
The all-cause mortality for a population can also be derived from the model, as $h_{\text{all}} + \frac{c}{S+C} h_f$. The quantity $\frac{c}{S+C} h_f$ is important in its own right, because for conditions that are unambiguously coded as the cause of death on death certificates, the population-level cause-specific mortality rate (i.e., the fraction of death certificates with this cause coded as the underlying cause of death) is approximately equal to this population-level excess mortality rate. Even for conditions where the death certificates are not likely to be coded to this cause for all deaths (e.g., diabetes), the population-level cause-specific mortality rate is a lower bound on $\frac{c}{S+C} h_f$, which can still provide useful information.

An important distinction between the hazards $h_{\text{with}}$ and $\frac{c}{S+C} h_f$ is the population to which they apply. The hazard $h_{\text{with}}$ applies only to the with-condition population, while $\frac{c}{S+C} h_f$ applies to the general population, including both those with and those without the condition. The hazard $h_{\text{with}}$ can be measured by following a cohort of individuals with the condition over time, while $\frac{c}{S+C} h_f$ can (sometimes) be measured by looking at deaths due to the condition in the general population.

Remission and duration studies, in which individuals with the disease are tracked over time to estimate how long the disease persists, provide yet another measurement that corresponds to parameters in this model (in the case of remission) or a quantity that can be derived from the model parameters (in the case of duration).

Representing mean duration of disease (i.e., the expected time an individual spends in compartment $C$ before leaving) is a relatively involved calculation, included here for completeness:

$$\text{duration}(a, t) = \int_{\tau=0}^{\infty} e^{-\left(h_r(a+\tau, t+\tau) + h_f(\tau+\tau) + h_m(a+\tau, t+\tau)\right) \tau} d\tau.$$  

This integral can be simplified considerably for certain special cases. For example, if the remission hazard is constant as a function of age and time, and the mortality hazards $h_m$ and $h_f$ are small, then the duration simplifies to $\int_{\tau=0}^{\tau} e^{-h_r(\tau+\tau)} d\tau = \frac{1}{h_r}$. This justifies the approximation that remission rate is nearly equal to 1 over duration in acute conditions.

The data available for these parameters varies widely among the diseases that have been analyzed in GBD 2010. Some diseases have many and more data, while others have only prevalence or only incidence and not much of that. These gaps and the bridges between what we know and what we want
to know will be explored in theory and practice in several later sections of this book.

With this detour through potentially available data in mind, it is now instructive to return to the more challenging elements of the compartmental model. Conceptually, it is excess mortality hazard $h_f$ that has proven hardest to explain and to understand. It can be interpreted as the difference between the mortality rates in the cases and controls in a cohort study. For this measurement of $h_f$ to be accurate, however, the study must avoid selection bias, which is quite a challenge in observational studies. Perhaps it would be clearer to focus on the with-condition mortality hazard, which can be measured directly in cohort study. With-condition mortality is not represented directly in the model but can be derived as $h_m = h_m + h_f$.

There are large differences in disease parameters such as incidence and prevalence as a function of age, and it is essential for a model to take this into account. Congenital abnormalities all have a birth prevalence, while important diseases such as dementia and Alzheimer’s disease have effectively zero prevalence in the young and dramatically increasing incidence and prevalence at older ages. Furthermore, the incidence and prevalence of disease, as well as the remission and excess mortality hazards, change over time due to shifts in population, changes in prevention or treatment, and changes in care. And the interdependence between these factors is complex but cannot be ignored: today’s population of 50-year-olds will be next year’s 51-year-olds.

For these reasons, a system of partial differential equations describing the change in the size of the compartments as a function of age and time provides a sufficiently rich theoretical framework for the model of process. In this formulation, the incidence, remission, without-condition mortality hazard, and excess mortality hazard are all functions of time and age, and the initial conditions for the stock of susceptible and with-condition populations at age 0 are functions of time.

### 7.3 Endemic equilibrium

The full model is often more complex than can be supported by available data, which is usually very sparse and very noisy. In order to simplify the modeling procedure and reduce the computational challenge of estimation, it is assumed that the disease parameters do not change substantially with
7.4. **FORWARD SIMULATION EXAMPLES**

respect to time (in other words, is in endemic equilibrium). To be precise, this is the assumption that the partial derivative of all stocks and all flows with respect to time is 0:

\[
\frac{\partial S}{\partial t} = \frac{\partial C}{\partial t} = \frac{\partial h_i}{\partial t} = \frac{\partial h_r}{\partial t} = \frac{\partial h_m}{\partial t} = \frac{\partial h_f}{\partial t} = 0.
\]

Although the primary use of this model is for inference of model parameters (sometimes called the “inverse problem”), it is instructive to apply it to the “forward problem” to show how hazards on incidence, remission, and excess mortality produce different prevalence curves. The next section explores this forward simulation through a series of examples to develop an intuition about how consistency forces interrelationships among prevalence, incidence, remission, and mortality.

### 7.4 Forward simulation examples

This section begins with a classic example from the history of generic disease modeling, the hypothetical example used in describing DisMod in the first GBD study.\(^\text{[14]}\) The incidence hazard increases linearly from age 0 to 100, and the remission and excess mortality hazards are constant with respect to age. Using all-cause mortality data from Southern sub-Saharan Africa and a birth prevalence of 0 to specify the forward simulation produces the outputs shown in figure 7.4.

When the age pattern of excess mortality changes to also linearly increase as a function of age, the prevalence curve becomes more clearly nonlinear, showing a condition that increases in prevalence quickly in young age groups but more slowly in older ages. Figure 7.5, panel (a), shows the results of this change.

Although the prevalence age pattern is largely determined by the remission, incidence, and mortality hazards, the birth prevalence can also change the shape dramatically. Figure 7.5, panel (b), shows the results of the same remission, incidence, and mortality hazards as in panel (a), but with a birth prevalence of 1.5%.

To summarize, this series of figures has shown the intuitive and less-than-intuitive way that the levels and age patterns of different epidemiological parameters must be interrelated to satisfy the fundamental equations of population health (when disease hazards for each age change negligibly slowly as a function of time).
Figure 7.4. Consistent disease parameters for a condition where incidence $h_i$ increases linearly as a function of age while remission $h_r$ and excess mortality $h_f$ hazards are constant. The background mortality $h_m$ has an age-specific hazard that follows all-cause mortality for females in the Southern sub-Saharan region in 1990. For a condition with prevalence of 0 at age 0, these hazards lead to a nearly linear increase as a function of age.

Figure 7.6 is designed to continue building familiarity with the features of consistent disease modeling by selecting age patterns for certain hazards to provide stylized examples similar to a variety of diseases. For example, for a disorder like dysthymia, for which there is a minimum duration of 2–3 years and low excess mortality, the consistency conditions produce a prevalence age pattern that looks like a smoothed version of the incidence age pattern, as shown in panel (a). For a congenital disorder, like Down syndrome, with birth prevalence, no incidence after birth, no remission, and substantial mortality, the consistent prevalence age pattern is shown in panel (b). For a disorder that affects the elderly, like Parkinson’s disease, the consistent age patterns for mortality, incidence, remission, and prevalence could look roughly like the age-specific rates shown in panel (c). And for disorders related to reproductive health, like prolapse, with zero excess mortality, incidence during ages 15–50, and remission that increases substantially at age 50, the consistent age patterns could look like those shown in panel (d). To conclude this series of plots, I have included an “incidence impulse response” example, showing the prevalence produced to be consistent with an incidence pattern that is nonzero for only a single age
Figure 7.5. Consistent disease parameters for a condition where incidence \( h_i \) and excess mortality \( h_f \) both increase linearly as a function of age while remission \( h_r \) is constant. The background mortality \( h_m \) has an age-specific hazard corresponding to females in the Southern sub-Saharan Africa region in 1990. Panel (a) shows that for a condition with prevalence of 0 at age 0, these hazards drive a prevalence age pattern that increases quickly in younger age groups and more slowly in older age groups. Panel (b) shows that for a condition with prevalence of 1.5% at age 0, these rates yield a prevalence age pattern that is highly nonlinear, dipping to a minimum of 1.3% at age 9 and then increasing back up to 1.8% at the oldest ages.

This is the content of panel (e).
Figure 7.6. Examples of different age-specific rates and the resulting prevalence curves.
7.5 Summary and future work

This chapter developed the compartmental model for disease moving through a population, which is the key to incorporating data from different epidemiological parameters (e.g., prevalence and incidence).

Future work must focus on the restriction to consider only the endemic equilibrium model of disease. This is justified in many cases, and required due to the sparse and noisy data in many more. However, there will be applications moving forward where disease rates are changing quickly over time and data is available to demonstrate this. In these cases, it will be necessary to solve the inverse problem without making the endemic equilibrium assumption, which will be a computational challenge.
Chapter 8

Numerical algorithms

Abraham D. Flaxman

Computational tractability has an important influence on model development, which often goes unacknowledged. The models that I fit are a compromise between models I would like to fit and the limitations of the algorithms and computing infrastructure available. This has always been the case, but modern algorithms and modern computing have shifted the balance tremendously.

In the days before digital computers, computational tractability meant that models had to be simple and computational methods elegant. In the 18th century, for example, an important challenge in predictive modeling was in navigation.\(^{54}\) Forecasting the path of stars allowed a ship to chart its course accordingly. The method of least squares, first published at the start of that century by Legendre, elegantly provided a solution.\(^{55}\) Using this method, mathematician-astronomers could plot the location of a celestial body at different time points, postulate a parametric model (e.g., that the body moves in a straight line), and then use the method of least squares to determine the parameters of the model that best fit the data. Why minimize the squared sum of the residuals? Why not minimize a different distance between the data and a proposed solution? Why not minimize the sum of the distance and the number of parameters in the model? The method of least squares has some appealing theoretical properties, since it is equivalent to finding parameters of maximum likelihood if the errors are normally distributed. But more importantly, minimizing the squared sum was a computational challenge well matched to the computational resource limitations of the 18th century. It is within reason to calculate the solution with pen and paper.
With the development of digital computation, more computationally intensive methods have become feasible. Topologist Stanislaw Ulam sparked the development of one such class of methods when he challenged himself to calculate the probability of winning in a variant of solitaire in the 1940s. An analytic solution was elusive, but a computationally intensive approximation method gave an approximate solution trivially, at least in theory. Ulam realized that it was more practical to repeat the solitaire game many times and count the number of successful plays than to estimate by pure combinatorial calculations. This approach has grown into the Monte Carlo method, a class of computational methods that rely on repeated random sampling to approximate calculations that are intractable or even impossible to calculate exactly.

The successors to the Monte Carlo algorithm make the Bayesian methods I use in integrative systems modeling possible. In Bayesian terms, the model of process and the model of data articulated in the previous chapters provide a prior distribution and likelihood. In equations, it is a simple application of Bayes’s formula to go from this to the posterior distribution. The exact computation of the distribution is intractable, however, and it is algorithms for sampling from the posterior distribution (or a close approximation thereof) that produce the parameter estimates for my models.

Bayesian methods were developed contemporaneously to the method of least squares but were limited in application before the development of Markov chain Monte Carlo (MCMC) algorithms and modern computers. Prior to these innovations, analysis was tractable for only a limited class of prior distributions and likelihoods. But with sufficient computing power, the posterior distribution can be sampled using Monte Carlo methods instead of being computed analytically. Monte Carlo methods can also be applied to integrate the posterior distribution to obtain, for instance, the posterior mean and variance. As computational resources to apply the approach to more complex problems have become more widely available, the approach has gained popularity.

The integrative systems model of disease in a population does not admit a closed-form representation for its posterior distribution. Instead, it relies on MCMC to draw samples of the model parameters from their posterior distribution. This, too, requires some care. The statistical computation tradition has put a lot of effort into deriving Gibbs samplers for specific Bayesian models, while theoretical computer scientists have focused on developing generic algorithms like the “Ball Walk” for sampling from convex
sets. The Metropolis-Hastings step method and the Adaptive Metropolis (AM) variant, in practice, provide acceptable performance without requiring burdensome derivation of customized Gibbs distributions. The MCMC algorithm benefits from wisely chosen initial values, and this seems to be particularly true when using MCMC with the AM step method in a large parameter space. Powell’s method optimizes a function of many variables without requiring derivatives, to find initial values for the model parameters for MCMC. Normal approximation at this initial value finds initial values for the variance-covariance matrices in the AM step method. Furthermore, an empirical Bayes approach separates the global model into submodels that can be fitted in parallel. The remainder of this chapter describes each aspect of the numerical algorithm in more detail.

8.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a class of Monte Carlo methods that obtains approximate solutions using a carefully designed Markov chain. A Markov chain is a stochastic process, or a sequence of random variables, such that the probability distribution of a random variable at one point in the sequence depends only on the random variable immediately before it in the sequence. If a Markov chain satisfies certain conditions (ergodicity), then it must tend toward a unique stationary distribution as the sequence continues. The key to using the MCMC algorithm for integrative systems modeling is constructing a Markov chain with the following three properties:

1. The stationary distribution of the chain is equal to the posterior distribution of the model.
2. Each step of the chain can be computed efficiently.
3. The chain converges to its stationary distribution in a reasonable number of steps.

A simple example can make this clearer. Suppose one wants to sample uniformly from the unit ball in $n$ dimensions, meaning the set of points \( \{ x \in \mathbb{R}^n : \|x\| \leq 1 \} \). The MCMC approach starts from any point in the ball, for example, the origin \( X_0 = (0, \ldots, 0) \), and generates successive points \( X_1, X_2, \ldots \), randomly, so that the points are Markovian, which is to say that the probability density of \( X_{i+1} \) is dependent only on the value of \( X_i \). There
is great art to designing the probability density that produces $X_{i+1}$. In this case, the Gibbs step is a simple one: choose an axis $e_i$ uniformly from the basis $\{e_1, e_2, \ldots, e_n\}$, and then choose $X_{i+1}$ uniformly from the interval given by the intersection of the ball with the line parallel to $e_i$ that passes through $X_i$.

To be precise, this Markov chain has transition probability density given by

$$p(X_{i+1} = x | X_i) = \begin{cases} \frac{1}{n} \cdot \frac{1}{2\sqrt{1 - \sum_{j \neq d} x_j^2}}, & \|x\| \leq 1 \text{ and } x_j = X_{i,j} \text{ for all } j \neq d; \\ 0, & \text{otherwise.} \end{cases}$$

When $n = 2$, it is possible to visualize this example in two dimensions, as shown in figure 8.1.

**Figure 8.1.** The results of drawing 100 samples from a ball in two dimensions with MCMC using the Gibbs step method. Although each sample is dependent on the previous samples, this dependence quickly decays, as shown by plotting the autocorrelation function of $X_1(t)$ and $X_2(t)$ in panels (d) and (e).

To see that the uniform distribution of the chain is stationary requires a small calculation. What is the probability density of a point $x \in \mathbb{R}^n$ after a single step of the chain? If $p(X_i = x) = 1/Z$ for all $x$, with the notation
\[ \ell_d = \sqrt{1 - \sum_{j \neq d} x_j^2}. \]

\[
p(X_{i+1} = x) = \int \int \cdots \int p(X_{i+1} = x|X_i = x') p(X_i = x') dx'_1 dx'_2 \cdots dx'_n \\
= \sum_{d=1}^{n} \int p(X_{i+1} = x|X_i = (x_1, \ldots, x_d, \ldots, x_n)) \\
\times p(X_i = (x_1, \ldots, x_d', \ldots, x_n)) dx'_d \\
= \sum_{d=1}^{n} \int_{-\ell_j}^{\ell_j} \frac{1}{n} \frac{1}{2\ell_d} Z dx'_d \\
= \frac{1}{Z}.
\]

Implementing each step of the chain requires only a way to choose numbers uniformly from the interval \([0, 1]\). This is not simple, but it is a basic primitive that randomized computation relies on; I use the Mersenne Twister pseudorandom number generator, a well-tested standard. To generate \(X_{i+1}\) from \(X_i\), the following suffices:

- Choose dimension \(d_i \in [n]\) uniformly at random.
- Choose sign \(s_i \in \{-1, 1\}\) uniformly at random.
- Choose fraction \(f_i \in [0, 1]\) uniformly at random.
- Set \(X_{i+1}\) equal to \(X_i\) for all coordinates besides \(d_i\) and let

\[ X_{i+1}(d_i) = s_i f_i \sqrt{1 - \sum_{j \neq d_i} x_j^2}. \]

Proving that Markov chains such as this one rapidly converge to their stationary distributions is a topic of current research in probability theory.

### 8.2 The Metropolis-Hastings step method

As mentioned above, my approach to parameter estimation with MCMC does not rely on deriving Gibbs step methods, which are often much more involved than the simple example in the previous section. I rely heavily
on the Metropolis-Hastings (MH) step method and an adaptive variant thereof.

In the context of Bayesian statistics, the MH algorithm is a technique used to sample from the posterior distribution when the posterior distribution cannot be easily sampled from directly. The algorithm generates the next position of its random walk in two steps. First, it makes a proposal by choosing from a proposal probability distribution, which depends on the current value of the walk. Second, it accepts or rejects this proposal with a probability carefully designed to yield the desired stationary distribution.

In the example from the previous section, uniform sampling from the unit ball, the proposal distribution could be a normal distribution centered at the current value, for example,

\[ P_i \sim \text{Normal}(X_i, \Sigma^2). \]

The MH rejection rule is based on the quantity \( p_i = \min\left(1, \frac{p(P_i)p(p_i|X_i)}{p(X_i)p(p_i|P_i)}\right) \), where \( p(\cdot) \) is the posterior probability density for value \( x \), and \( p'(p, x) \) is the probability density of proposing \( p \) when the chain has value \( x \). The rejection rule is

\[
X_{i+1} = \begin{cases} 
P_i, & \text{with probability } p_i; \\
X_i, & \text{with probability } 1 - p_i. 
\end{cases}
\]

When sampling from the unit ball with the symmetric proposal distribution above, this simplifies to

\[
X_{i+1} = \begin{cases} 
P_i, & \text{if } \|P_i\| \leq 1; \\
X_i, & \text{otherwise}. 
\end{cases}
\]

Making the example from the previous section only a little bit more complicated demonstrates both the utility and the challenges of the MH step method for MCMC. Instead of sampling from the unit ball, now I will consider sampling uniformly from an ellipsoid in \( n \) dimensions, \( \{x \in \mathbb{R}^n : x^T \Lambda x \leq 1\} \). In this case, the Gibbs step method requires solving a system of equations at each step to determine the limits of the ellipse along the selected dimension. The MH step method requires only testing whether the proposed point is in the ellipse. On the other hand, the Gibbs step method always moves to a new point in the sample space, while MH sometimes
rejects the proposal and stays at the same point for multiple steps. In either case, if the ellipse is long and skinny, it will slow the chain. The Gibbs steps will not move very far much of the time, while the MH steps will often not move at all.

When \( n = 2 \), it is possible to visualize this example in two dimensions, and figure 8.2 shows the results for an ellipse with width three times its height.

![Figure 8.2](image)

**Figure 8.2.** The results of drawing 100 samples from an ellipse in two dimensions with MCMC using the MH step method. Each sample is dependent on the previous samples, and because of the shape of the ellipse, the MH proposals are often infeasible, so the dependence does not decay rapidly, as shown by plotting the autocorrelation function of \( X_1(t) \) and \( X_2(t) \) in panels (d) and (e).

### 8.3 The Adaptive Metropolis step method

The Adaptive Metropolis (AM) step method extends the MH step method by adaptively adjusting the variance-covariance matrix for the proposal distribution based on the acceptance rate of the proposals. Because the proposal acceptance rate is so important to algorithmic efficiency, a line of research has considered adaptive approaches to proposal distribution selection.

One popular adaptive approach begins with a simplified version of the MH step method above (often called the Metropolis step method), where a
propose is generated at each step

\[ P_i \sim \text{Normal} \left( X_i, \Sigma_i^2 \right), \]

and then accepted or rejected with probability

\[ p_i = \min \left( 1, \frac{p(\mathcal{P}_i)}{p(X_i)} \right) \]

This is a simplification of the MH step method, because the terms about the transition probability are not included in the proposal acceptance probability. But it has a subtle complexification of the MH step method as well, because the proposal distribution covariance matrix \( \Sigma_i \) now changes as the chain progresses. The goal is to change it in a way that adapts to the distribution from which it is being used to sample.

The adaptive values of \( \Sigma_i \) that I have used follow those implemented in the PyMC software package, \( ^{64} \)

\[ \Sigma_i = \begin{cases} \Sigma_0, & i \leq i_0; \\ s_n \left[ \text{cov}(X_0, \ldots, X_i) + \epsilon I_n \right], & i > i_0. \end{cases} \]

The initial value for the covariance matrix, \( \Sigma_0 \), has a large influence on the time the MCMC algorithm takes to converge. For the additional parameters, I have used the PyMC default values, where in an \( n \)-dimensional sample space, I have \( s_n = (2.4)^2/n, \epsilon = 10^{-5} \), and \( I_n \) (the \( n \)-dimensional identity matrix).

For additional computational speedup, I have followed the PyMC modification of the original AM step method and updated \( \Sigma_i \) only every 100 or 1000 steps. Furthermore, if few proposals are ever accepted, then I decrease the variance of the proposal distribution by a constant factor.

When \( n = 2 \), the results of this step method can be visualized in two dimensions, and figure 8.3 shows the results of the AM stepper after 20,000 iterations have been run to allow the covariance matrix to adapt to the posterior distribution.

8.4 Convergence of the MCMC algorithm

The chief pitfall of the MCMC algorithm is nonconvergence. A line of theoretical research in computer science is devoted to identifying classes of distributions and classes of step methods for which MCMC can be proven to converge rapidly. \( ^{69,70,71,72} \) Unfortunately, this work has recently developed
8.4. CONVERGENCE OF THE MCMC ALGORITHM

Figure 8.3. The results of drawing 100 samples from an ellipse in two dimensions with MCMC using the AM step method, after 20,000 iterations of “burn-in.” Each sample is dependent on the previous samples, but after a period of exploration, the AM proposals become aligned with the axes of the ellipse, and the dependence decays more rapidly than in the MH approach in figure 8.2.

strong evidence against the possibility of automatically detecting convergence of MCMC in general. This work has developed largely separately from that reported in the Bayesian computing literature, where posterior estimates generated by MCMC sampling are used frequently in important settings. In the applied statistical literature, MCMC is too convenient not to be used, and since nonconvergence is such an important pitfall, a vast array of heuristic checks for convergence have developed over the years. A recent survey compared and contrasted many of these.

Essential to using MCMC for statistical estimation is reproducibility, and here is where nonconvergence wreaks its havoc. MCMC computation is randomized computation, meaning that the same algorithm run on the same data twice will give slightly different answers. This is fine, as long as the variation between successive runs can be controlled. When MCMC has not converged, the difference between two runs cannot be predicted, which means that the results will not be reproducible. This must be prevented for MCMC computation to be useful.

One heuristic to identify nonconvergence that has been particularly useful in my work is to visually inspect the autocorrelation plot for each di-
mension of the sample space. This autocorrelation function is defined by

\[ \text{acf}(\tau) = \frac{\mathbb{E}[(X_i - \mu)(X_{i+\tau} - \mu)]}{\sigma^2}, \]

where \( \mu \) and \( \sigma \) are the mean and standard deviation of the posterior distribution. The autocorrelation plot for independent samples is a delta function, and the rate of decay of the autocorrelation plot gives some indication of how close the samples are to being uncorrelated. Panels (d) and (e) in figures 8.1, 8.2, and 8.3 show autocorrelation plots in a provably rapidly mixing chain (figure 8.1), in a clearly nonconvergent case (figure 8.2), and in a marginal case that I would run longer to be sure (figure 8.3).

There are three general approaches to improve the convergence of MCMC computation. The first and simplest approach is to run the chain longer and thus take more samples. In the long run, the algorithm will succeed, and the only question is whether the program can run for long enough in the time available for the analysis. The second approach is to use a more appropriate step method—for example, by using AM steps instead of the MH steps in the ellipse example above. Other, more complicated step methods can also be used, and the development of new and improved step methods is an active area of research. The third approach is to use better initial values for the MCMC, which includes both starting from a likely point in the posterior distribution and, in the case of AM and other advanced step methods, initializing the step method parameters wisely as well.

The wise selection of initial values is the topic of the next section.

### 8.5 Initial values for MCMC

Past research has found that the choice of initial values for MCMC can dramatically affect the time necessary for convergence. In theoretical work, a “warm-start” increases performance by selecting initial values that are near the target values. Ideally, warm-start values are from a distribution where the density at any point is at most twice the density of the target distribution. In practice, a number of different methods have been proposed for generating initial values for MCMC samples. The simplest approach is to choose initial values from the prior distribution. In my experience, this is not as stable as choosing initial values based on the results of a local optimization procedure that finds an initial value that approximately maximizes the posterior distribution.
8.6 Empirical Bayesian priors to borrow strength between regions

When making estimates for all 21 geographic regions in the GBD Study 2010, the available data are often very sparse. I have often used an approach to borrow strength between regions. It is a two-stage approach that can be characterized as an empirical Bayesian technique.

The mechanics of this approach are simple. First, I fit a model to the available data at the global level. Depending on the data available, this is often either an inconsistent model, using the random-effect age-integrating negative-binomial spline model from chapters 3–6, or a consistent model linking random-effect age-integrating negative-binomial spline models for different epidemiological parameters through the system of differential equations in section 7.2. This model is used to make estimates for all regions of the world.

The second stage of the empirical Bayes approach is to use the first-stage predictions for a region as priors and to fit the model again, now including the empirical priors, with only the data relevant to a particular region, sex, and year.

8.7 Summary and future work

The MCMC algorithm, with the AM step method, has been an enabler for this entire approach. Without free/libre open-source software for implementing AM/MCMC, this project would not have been possible. Initial values from optimization of the posterior distribution, and the empirical prior approach to decompose the regional estimation tasks into independent calculations were also crucial to finding reasonable answers in the available computation time.
However, this MCMC-based computation is not the only algorithmic approach. Message passing algorithms have proven themselves quite successful in related computational challenges. Nonlinear optimization is another promising approach, especially combined with bootstrap method for estimating uncertainty. As computational resources continue to evolve, and new computational algorithms are developed, the possibility of incorporating new innovations into faster and more accurate estimates should be continually explored.
Part II

Applications
Chapter 9

Knot selection in spline models: cocaine dependence

Yong Yi Lee, Theo Vos, Abraham D. Flaxman, Jed Blore, and Louisa Degenhardt

For many conditions prevalence varies substantially as a function of age. Other epidemiological parameters, such as incidence and excess mortality hazards, often have important age patterns as well. The spline models introduced in chapter 3 provide a flexible framework for representing this age dependence. However, some important modeling decisions are necessary. The following examples from estimating the age-specific prevalence of cocaine dependence illustrate the importance of choosing knot locations and smoothing levels appropriately in a setting where the data speak relatively precisely about the level and age pattern of the condition.

The American Psychiatric Association’s *Diagnostic and Statistical Manual of Mental Disorders, Version IV, Text Revision (DSM-IV-TR)* recognizes cocaine dependence as fulfilling 3 or more of the following 7 dependence criteria during any time in the same 12-month period: (86,87)

- tolerance to effects of cocaine (typically assessed by whether the same amount of cocaine has less effect or whether greater amounts are required to obtain the desired effect);
- withdrawal symptoms after use ceases;
- usage over longer period or in larger quantities than intended;
- persistent desire or unsuccessful efforts to control cocaine use;
CHAPTER 9. KNOT SELECTION IN SPLINE MODELS: COCAINE DEPENDENCE

- substantial time spent in obtaining, using, or recovering from effects of cocaine;
- reduction of important social, occupational, or recreational activities because of cocaine use;
- continued use despite knowledge of physiological or psychological problems induced by cocaine use.

Despite a large body of data on cocaine use, there are comparatively few data available on the descriptive epidemiology of cocaine dependence. Systematic review for cocaine dependence identified 28 prevalence data points, covering 3 GBD 2010 regions. For this example, we have restricted our attention to data from the USA (figure 9.1).

As discussed in chapter 3, we model age-specific hazards with spline models. In this case, the spline model takes the form of a continuous, piecewise linear function, with selected “knots” where the function is non-linear. These knots partition the age range into intervals, and the choice of knots can be influential for the resulting estimates. In a setting where data are not sparse and noisy, estimates will not be very sensitive to the choice of knots. However, when working with sparse and noisy data, the number and location of knots are important decisions, as they can influence the model results substantially. Thus, the number of knots and locations should be chosen a priori using expert knowledge concerning the disease being modeled. It is also important to consider additional knots and alternative configurations of knots as a sensitivity analysis.

To demonstrate the importance of the number of knots in a spline model, we compare three models with differing numbers of knots in figure 9.2. The 7-knot model for cocaine dependence has knot set \{15, 20, 25, 30, 40, 50, 65\}, based on the theory that prevalence is 0 in childhood, changes rapidly during early adulthood, and then changes less rapidly at older ages. The age pattern of prevalence estimated with this model is shown as a solid line with the 95% highest posterior density (HPD) shown as thin solid lines.

Another model we consider is a 4-knot model, which has knots set \{15, 25, 40, 65\}. The estimates from this model, shown as a dotted line, demonstrate how, seemingly paradoxically, fewer knots can lead to more extreme estimates for certain ages.

We also fitted a model with 11 knots, using knots spaced every 5 years from ages 15 to 65. The estimates from this model are shown as a dashed
Figure 9.1. Prevalence data for cocaine dependence in the USA. Each horizontal bar represents a single data point extracted in systematic review. The left and right endpoints indicate the start and end ages of the age interval for a data point, while the level of prevalence is represented by the distance of the bar above the x-axis.

When data are sparse, adding additional knots allows for estimates that follow the vagaries of the data more closely, which may not be desired. Smoothing priors for the penalized spline model can address this.

The penalized spline model introduces an additional term to the model prior to encode the belief that the age pattern is not too wiggly. With the judicious choice of the smoothness hyperparameter, the model can include more knots without using them to chase the noise around in the noisy data. The effects of four values of the smoothing parameter are shown in figure 9.3. The smaller the parameter, the smoother the estimated age
pattern and, hence, the less influential the position of the knots. However, too much smoothing leads to overcompression of the prevalence estimates, resulting in estimates that are not representative of the data. If there were enough time and data, it would be ideal to compare out-of-sample predictive validity for a range of knots and smoothing parameters, although in-sample goodness-of-fit statistics such as the Bayesian information criteria (BIC) or deviance information criteria (DIC).

As shown by the figures, knot location and smoothing hyperparameter selection can be influential parts of the model. From the sensitivity analysis in figure 9.2, it appears that there is nothing to gain from adding additional knots to the model, although it is possible to do so if an appropriate smoothing parameter is included also. In the next chapter, we will consider a case
Figure 9.3. Prevalence estimates from the model with knots spaced every 5 years using a penalized spline model with a smoothing parameter $\sigma$.

where the data do not have such a clear story to tell about the age pattern.
Chapter 10

Unclear age pattern, requiring expert priors: premenstrual syndrome

Hannah M. Peterson, Yong Yi Lee, Theo Vos, and Abraham D. Flaxman

Epidemiological data without clear age patterns are a reoccurring theme in GBD 2010. Unclear age patterns make expert priors essential in the modeling process. However, such cases are very sensitive to the choice of prior assumptions, as shown in the following example of premenstrual syndrome (PMS) in Western Europe.

PMS is a common cyclic disorder that affects women of reproductive years during the period between ovulation and the onset of menses. More than 200 behavioral, psychological, and physical symptoms have been associated with PMS, the most common being irritability, tension, depression, bloating, weight gain, and food cravings. The exact causes of PMS are unknown, with no consistent treatment option available.

A systematic review of the descriptive epidemiology of PMS yielded 74 prevalence data points, of which 18 were from Western Europe. As seen from figure 10.1, the data are noisy, with overlapping and heterogeneous age groups that show no clear age pattern.

In the absence of clear age patterns in the systematic review data points, modeling decisions about knot location, age pattern levels, and direction of age pattern trends have substantial influence on the estimates of disease prevalence. These decisions can also have unintended consequences, as discussed in chapter 3. To illustrate the effects, we fitted models with a variety
CHAPTER 10. UNCLEAR AGE PATTERN, REQUIRING EXPERT PRIORS: PREMENSTRUAL SYNDROME

Figure 10.1. Prevalence data for women with PMS in Western Europe.

of choices about knot location, level values outside the measured age intervals, and direction of age pattern trends. Conducting a sensitivity analysis like this is important when modeling sparse and noisy data. The sensitivity analysis allows the modeler to understand the range of answers for the given assumptions and provides a long term guide for identifying areas of future research to collect more precise estimates.

As the prevalence data plotted in figure 10.1 show, systematic review collected no data on population prevalence for women younger than age 15 or older than age 50. Since PMS is a disorder related to the female reproductive cycle, it follows that data outside this age range are not present for biological reasons. However, this information is not part of the spline model unless the modeler explicitly includes it. If no priors are included to inform estimates in the young and old, then they are extrapolated from the levels
Figure 10.2. Four estimates of age-specific PMS prevalence with different priors on prevalence in young and old ages. Systematic review produced sparse and noisy data, shown here for Western Europe. (a) Without a level prior to inform the model that prevalence data are not present outside ages 15–50 for biological reasons, estimates outside the ages measured are extrapolated from inside. Restricting prevalence to 0 changes the prevalence estimates substantially. (b) The effect of assuming $p(a) = 0$, for $a < 15$, (c) the effect of assuming $p(a) = 0$, for $a > 50$, and (d) the effect of assuming $p(a) = 0$, for $a < 15$ and $a > 50$ all change the estimates inside and outside the observed data ages.

where there are data, as seen in figure 10.2. The sensitivity analysis shown in figure 10.2 is to highlight the effect of modeling assumptions; prevalence at ages younger than 15 and greater than 50 is biologically implausible and expert knowledge is needed to inform the model that cases are not expected outside the age range where they have been measured.
As described in section 3.1, we model age-specific hazards with splines, using knots to partition the age range into intervals. Models with ample data and clear age patterns are not very sensitive to knot choice. However, with sparse and noisy data without a clear age pattern, the number and location of knots can influence the model results substantially. We explored this by fitting models with a variety of knots to the PMS data set, as seen in figure 10.3. Choosing the number and location of knots a priori using expert knowledge allows the user to determine critical features of the model in a principled way.

Another common prior for age patterns is the belief that the epidemiological parameter increases or decreases over a certain age range. As seen in figure 10.4, priors on monotonicity between the critical ages of 25 and 40 have a large effect on the prevalence estimate for Western Europe.

Knot selection and priors on level and monotonicity play an important role in the modeling process and in the sensitivity analysis. However, when the data are not sufficient to understand the age pattern, the model compensates by producing estimates with large uncertainty, as seen in figure 10.5. This estimate comes from a model with knots at \{0, 15, 20, 30, 40, 50, 100\}, no prior on monotonicity, and a prior on level to restrict prevalence to be 0 outside the age range 15–50.

This example has identified an area of future research. The model does not have enough data to inform an age pattern because the descriptive epidemiology of PMS is quite uncertain: some studies say almost all women experience it and some studies say none do. In such cases, making the most informed decisions possible (such as restricting the model to ages 15–50 for biological reasons) and accepting a large uncertainty interval reveal the truth: we just don’t know.
Figure 10.3. Estimate of age-specific PMS prevalence for spline models with a variety of knots. All panels have knots at \{0, 15, 50, 100\} and vary the number and location of knots between the ages of 15 and 50 to show the sensitivity of knot selection in sparse and noisy data without a clear age pattern. (a) With 1 additional knot, the placement at age 20, 32, or 45 gives markedly different estimates of PMS prevalence in Western Europe. (b) With 2 knots at \{27, 38\}, \{20, 45\}, or \{30, 35\}, the differences are also clear and predictable. (c) With 3 knots at locations \{23, 32, 41\}, \{18, 32, 47\}, or \{29, 32, 35\}, it appears that the data are too sparse and noisy to support a consistent age pattern.
CHAPTER 10. UNCLEAR AGE PATTERN, REQUIRING EXPERT PRIORS: PREMENSTRUAL SYNDROME

Figure 10.4. Estimates of age-specific PMS prevalence for spline models with a variety of monotonicity priors. Between the ages of 25 and 40, the prior on monotonicity makes a large impact on the prevalence estimates for women in Western Europe with PMS.
Figure 10.5. Prevalence estimates for women in Western Europe with PMS.
Chapter 11

Empirical priors: pancreatitis

David Chou, Hannah M. Peterson, Abraham D. Flaxman, Christopher J. L. Murray, and Mohsen Nagavi

Systematic review for GBD 2010 often found a few regions for which detailed data on the age pattern of disease were available, but many more regions for which cases were reported with much sparser age specificity. Hierarchical modeling using an empirical Bayesian prior is our way to conduct partial pooling and borrow strength from the regions with age-specific data to produce estimates of age patterns in regions where few or no age-specific data are available. This chapter demonstrates the results of partial pooling at the regional level, where country-to-country variation is quite pronounced, by examining the estimation of age-specific pancreatitis incidence in Western Europe.

Pancreatitis is the inflammation of the pancreas, most commonly caused by alcohol or gallstones. In most cases, the disease resolves itself and there is no need for treatment. However, some acute cases develop pancreatic necrosis and systemic organ failure. These complications require immediate treatment and have a high mortality risk.

Data from systematic review yielded 3950 incidence data points, 1053 of which were from Western Europe and constitute the example in this chapter. As shown in figure 11.1, the data from Western Europe are very heterogeneous and the pooled estimates do not reflect the data the best, especially between the ages of 25–60.

Closer investigation in figure 11.2 shows that there is considerable heterogeneity caused by different age patterns between countries. The pooled estimate in figure 11.1 (dashed line, figure 11.1) does not capture this vari-
Figure 11.1. Pancreatitis incidence data with pooled estimates for males and females in Western Europe in 2005.

Empirical priors have the benefit of coping with large differences in age-specific rates and borrowing strength between countries. The empirical prior is the estimate from the pooled regional data from figure 11.1. A posterior distribution for each country can be estimated with this empirical prior. With lots of data, the empirical prior is irrelevant, as the data informs the posterior estimate. When there are no data, the posterior estimate follows the empirical prior with large uncertainty intervals since the countries with data show a lot of country-to-country variation (this large uncertainty is also the reason why the means of the empirical prior and posterior estimates for Germany in panel (d) do not match precisely). With
some data, the posterior finds the proper balance between the data and the empirical prior.

Figure 11.2. Comparison of pancreatitis incidence estimates for males in 2005 for (a) Finland, (b) the Netherlands, (c) Cyprus, and (d) Germany. The estimated incidence using pooled data from figure [11.1] was applied as an empirical prior to the sex-specific incidence to improve estimates.

Empirical priors provide data-derived relationships to guide the modeling process. In cases where the data are less clear, empirical priors provide a principled and computationally tractable way to borrow strength between regions for estimation.
Chapter 12

Overlapping, heterogeneous age groups: atrial fibrillation

Mohammad H. Forouzanfar, Abraham D. Flaxman, Hannah M. Peterson, Mohsen Nagavi, and Sumeet Chugh

Like many conditions analyzed in GBD 2010, atrial fibrillation (AF) has no standard set of age groups for reporting. The meta-analysis of the data collected in systematic review must address these heterogeneous age groups in some way. AF provides a prototypical example of this, one where the results of the choice to use an age-standardizing model can be compared with those of other possible choices. This chapter compares the estimates produced for AF prevalence and incidence using an age-standardizing model with those from a midpoint model.

AF is the most common type of cardiac arrhythmia. Chaotic and irregular heart rhythms originating in the atria cause poor blood flow to the body. The duration of AF episodes varies greatly. Paroxysmal AF is occasional, with attacks lasting a few minutes or hours, whereas persistent AF and permanent AF are chronic, continuing for days with or without self-termination. Symptoms include heart palpitations, lack of energy, dizziness, shortness of breath, and chest discomfort, although some cases of AF are symptomless. AF may occur at any age, with increasing risk for older ages, and is uncommon in children. Other heart diseases tend to be the underlying cause of AF. AF is associated with coronary heart disease, hypertensive heart disease, valvular heart disease, heart failure, cardiomyopathy, obesity, and metabolic disorders such as diabetes and hyperthyroidism.

The GBD Study 2010 defines AF as a patient having at least one episode confirmed by a physician. The systematic review of AF collected 3942
data points, of which 247 were from countries in Western Europe. We will consider only the Western European data in this chapter. We have 20 data points on disease incidence and 147 on prevalence. As seen from figure 12.1, AF has heterogeneous and overlapping age groups. Without access to the microdata needed to re-create homogeneous age groups, combining all these data must rely on age-group modeling, as described in chapter 5.

As discussed in section 5.2, the simplest approach to modeling heterogeneous age groups is to apply each age-specific rate measurement to the midpoint of the age interval. Another solution to the heterogeneous age groups is to use age standardizing (section 5.5). Age standardizing adds age weights to the age-specific rate according to population structure. The age-standardizing model uses a common age pattern for all studies so that the age weights are the same for all country-years, as discussed in more detail in section 5.5.

As the prevalence estimates in figure 12.2 show, model choice changes the estimates. In estimates before age 80, differences are minimal, but in estimates for older ages, where the data are sparser and noisier, the differences are substantial.

Without additional information, one cannot say which model is preferred. Further investigation with incidence does not provide much insight. Figure 12.3 shows that, like the prevalence estimates, the incidence esti-
Figure 12.2. Comparison of estimates of prevalence of AF for Western European males in 1990: (a) data and estimates for the age-standardizing model; (b) data and estimates for the midpoint model.

The compartmental model estimates for incidence in figure 12.4 are very different from the spline model estimates. Unlike the spline models, the compartmental model estimates for incidence do not go through all the data. This is because the compartmental model requires internal consistency; that is, for every prevalence case there must be a matching incidence event.
CHAPTER 12. OVERLAPPING, HETEROGENEOUS AGE GROUPS:  
ATRIAL FIBRILLATION

The compartmental model shows that these levels of prevalence cannot be achieved with the levels of incidence the data show.

Figure 12.5 compares the prevalence and incidence estimates from the age-standardizing compartmental model with those from the midpoint compartmental model. As with the spline models, the estimates differ substantially only in the oldest ages.

The choice of age-group model has implications for disease estimates. Estimated age pattern, trends, and levels can differ between the midpoint and age-standardizing models. The age-standardizing model is a unique feature of the approach developed in this book, and allows more appropriate use of systematic review data than simply applying the measurement to the midpoint of the age interval.

Figure 12.3. Comparison of estimates of incidence of AF for Western European males in 1990: (a) data and estimates for the age-standardizing model; (b) data and estimates for the midpoint model.
Figure 12.4. Estimates of AF in Western European males in 1990 using an age-standardizing compartmental model for (a) prevalence and (b) incidence.

Figure 12.5. Comparison of the estimated (a) prevalence and (b) incidence in Western European males with AF in 1990 using age standardizing and midpoint compartmental models.
Chapter 13

Dealing with geographical variation: hepatitis C virus infection

Abraham D. Flaxman, Khayriyyah M. Hanafiah, Justina Groeger, Hannah M. Peterson, and Steven T. Wiersma

One challenge in global disease modeling for descriptive epidemiological estimation is properly reflecting the true regional variation in disease epidemiology. While some diseases are relatively consistent in their levels and age patterns from region to region, others vary a great deal. The most extreme examples of the latter type are focal diseases that are present only in certain regions, but the hardest to model are diseases that are present globally, but to greater and lesser degree. Hepatitis C virus (HCV) is an example of such a disease, which we examine in this chapter. In the absence of any strongly predictive covariates, we use hierarchical random effects to model this regional variation.

Hepatitis C is a disease caused by the viral infection of HCV, an RNA virus in the Flaviviridae family that predominately attacks the liver. In a small portion of acute cases, the body can eliminate the virus; however, the majority of acute cases develop into chronic infections. Chronic infections cause liver damage and may develop into end-stage liver disease, or cirrhosis. Few of those persons who are chronically infected experience symptoms, and only one-third of acute cases develop jaundice or other symptoms. Chronic symptoms are nonspecific, intermittent, and mild, with the most common symptom being fatigue. Common symptoms for severe and advanced disease stages include nausea, dark urine, and jaundice. Since
CHAPTER 13. DEALING WITH GEOGRAPHICAL VARIATION: HEPATITIS C VIRUS INFECTION

HCV infections are often asymptomatic, diagnosis usually requires laboratory testing for both hepatitis antibodies (anti-HCV) which, when positive, indicates past or current infection and the HCV nucleic acid (HCV RNA) which, when positive, indicated current infection. There is no vaccine that protects against HCV infection, but new treatments have been reported to clear the infection and prevent advanced liver disease. (99,100,101)

Compared with other countries in the North Africa and Middle East region, Egypt has a high prevalence of HCV infection in the general population. In an attempt to treat endemic schistosomiasis, a common parasitic worm that affects the urinary tract, gut, and liver, the Egyptian Ministry of Health launched widespread injection-based treatment from 1950 to 1980. While there were improvements in schistosomiasis-induced mortality, recycled needles and poor needle sterilization used to deliver these medicines inadvertently infected many with HCV. (102,103,104) The spatial variations of HCV infection in North Africa and the Middle East provides a striking example for hierarchical random-effects modeling.

Random-effects modeling detects systematic differences among different spatial units of data. The spatial hierarchy in the GBD Study 2010 uses countries nested in regions nested in superregions. There are 21 regions defined by demographic and epidemiological similarities, which are further clustered into 7 superregions (see appendix A).

The analysis of HCV infection uses data on the prevalence of persons who have antibodies to HCV (anti-HCV). Incomplete data or data from high-risk populations, such as injection drug users or paid blood donors, were excluded. Figure 13.1 shows data collected in systematic review for two countries in the North Africa and Middle East region, Egypt and Jordan. Notice that for some age groups, anti-HCV prevalence measurements in Egypt are more than 40 times higher than the corresponding measurements in Jordan.

We used the age-standardizing, generalized negative-binomial spline model with hierarchical random effects to estimate anti-HCV prevalence. The hierarchical random effects allow the model to capture variation within the North Africa and Middle East region. Table 13.1 shows that Egypt has significantly higher prevalence than the other countries in the region. Figure 13.2 confirms this, as the prevalence estimate for Egypt is much above the regional average.

In addition to hierarchical random effects, the negative-binomial rate model includes a parameter that estimates the amount of nonsampling er-
Figure 13.1. Prevalence data from systematic review of anti-HCV in (a) Egypt and (b) Jordan.

Table 13.1. Estimates of the country-level random effects for anti-HCV prevalence (an intercept shift in log-space) from the random-effects model for the countries in the North Africa and Middle East region

<table>
<thead>
<tr>
<th>Country</th>
<th>Posterior Mean</th>
<th>Lower 95% HPD</th>
<th>Upper 95% HPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt</td>
<td>1.87</td>
<td>1.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Jordan</td>
<td>-0.59</td>
<td>-1.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>-0.77</td>
<td>-1.2</td>
<td>-0.4</td>
</tr>
<tr>
<td>Iraq</td>
<td>0.07</td>
<td>-0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Iran</td>
<td>0.02</td>
<td>-0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Yemen</td>
<td>0.04</td>
<td>-0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Turkey</td>
<td>-0.31</td>
<td>-0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Syria</td>
<td>-0.13</td>
<td>-0.7</td>
<td>0.4</td>
</tr>
<tr>
<td>Tunisia</td>
<td>-0.19</td>
<td>-0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

error, parametrized as the overdispersion term $\delta$, described in chapter 2. In such noisy data, weakly informative priors on $\delta$ help with convergence and inform the posterior estimates about beliefs regarding data heterogeneity. This allows the model to incorporate expert beliefs about how much of the country-to-country variation is true variation and how much is nonsampling error, which can be important in the presence of sparse and noisy data.
We compared three alternative priors on the negative-binomial model dispersion parameter $\delta$, corresponding to “very,” “moderately,” or “slightly” overdispersed. The natural logarithm of $\delta$ is uniformly distributed between its lower and upper bounds. Intended as a weakly informative prior, the bounds of the categories of $\delta$ overlap, so that the bounds of “very,” “moderately,” and “slightly” are $[1, 9]$, $[3, 27]$, and $[9, 81]$, respectively.

In this example, the effects of priors on the overdispersion of $\delta$ are seen in the posterior estimates at the country level as shown in figure [13.3]. Random-effects modeling detects within-sample variation and true variation that cannot be explained by a covariate. Therefore, a change in the prior on global heterogeneity changes the level of variation and thus the size of...
the random effect. As seen in figure [13.3] when the prior on overdispersion is “very,” the estimates are more compressed than those with a prior of “slightly.”

![Figure 13.3](image)

**Figure 13.3.** The intercept shift of anti-HCV prevalence for men in 1990 in log-space with different priors on global heterogeneity, $\delta$. Four levels (global, superregion, region, country) were used in the hierarchical random-effects spline model.

Another way to view compressed estimates is by looking at the age-standardized prevalence in table [13.2]. As heterogeneity increases from “slightly” to “very,” country estimates are compressed toward the regional mean.

Hierarchical random effects and the overdispersion parameter $\delta$ allow the model to distinguish between true country-to-country variation and nonsampling error. Weakly informative priors on $\delta$ incorporate the modeler’s beliefs about data heterogeneity, while the hierarchical random effects
CHAPTER 13. DEALING WITH GEOGRAPHICAL VARIATION:  
HEPATITIS C VIRUS INFECTION

Table 13.2. Anti-HCV age-standardized prevalence estimates from a hierarchical random-effects spline model with differing priors on global heterogeneity

<table>
<thead>
<tr>
<th>Geographic Area</th>
<th>Heterogeneity</th>
<th>Posterior Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Africa and Middle East</td>
<td>δ \sim \text{Uniform}(9, 81)</td>
<td>0.048</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>δ \sim \text{Uniform}(3, 27)</td>
<td>0.049</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>δ \sim \text{Uniform}(1, 9)</td>
<td>0.045</td>
<td>0.005</td>
</tr>
<tr>
<td>Jordan</td>
<td>δ \sim \text{Uniform}(9, 81)</td>
<td>0.007</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>δ \sim \text{Uniform}(3, 27)</td>
<td>0.010</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>δ \sim \text{Uniform}(1, 9)</td>
<td>0.020</td>
<td>0.007</td>
</tr>
<tr>
<td>Egypt</td>
<td>δ \sim \text{Uniform}(9, 81)</td>
<td>0.188</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>δ \sim \text{Uniform}(3, 27)</td>
<td>0.179</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>δ \sim \text{Uniform}(1, 9)</td>
<td>0.137</td>
<td>0.019</td>
</tr>
</tbody>
</table>

provide a way to model regional variation. When a predictive covariate is available, it can be used instead of random effects to explain the country-to-country variation (as chapter 15 will demonstrate), but in the absence of this, geographic random effects provide a mechanism to model the level of unexplained but true variation between different areas.
Chapter 14

Cross-walking with fixed effects: anxiety disorders

Amanda Baxter, Jed Blore, Abraham D. Flaxman, Theo Vos, and Harvey Whiteford

The data collected in systematic review often contain a variety of different study types or diagnostic criteria, which create systematic biases in the measured data. An extreme example was found in the systematic review of diabetes prevalence, where there were 18 variants of diagnostic criteria. The systematic review of anxiety disorders provides a simpler example, which is the focus of this chapter. This systematic review collected studies that used a handful of different recall periods to ask about the presence of the disorders. The quantity of interest for estimation was point prevalence, the proportion of the population with the condition at an instant in time. We used a fixed-effect model to adjust for the bias introduced by studies that measured period (e.g., past-year) prevalence, since these studies also provide valuable information on the descriptive epidemiology of the condition. This bias adjustment by fixed-effect modeling is also called a “cross-walk.”

Anxiety disorders include at least 8 separate conditions each characterized by prominent anxiety at a level that interferes with daily life. Not all anxiety disorders manifest in similar ways. While generalized anxiety disorder is typically marked by persistent worry, panic disorder is usually characterized by intense fear for discrete periods of time. As there is a lot of comorbidity between individual anxiety disorders, anxiety disorders were modeled together as a single condition in GBD 2010.

Anxiety disorders do not have a consistent recall period for the measurement of epidemiological rates. Therefore, the data from systematic
review include studies with measurements of point prevalence and period prevalence (i.e., 6-month or past-year prevalence). The analysis excludes lifetime prevalence measurements because such estimates are particularly prone to recall bias. Due to the nonnegligible remission rate for anxiety disorders, period prevalence is typically higher than point prevalence, as seen in figure 14.1.

Figure 14.1. A comparison of (a) point and (b) period prevalence data for anxiety disorders in Australasian females, collected in a systematic review for 2000–2008.

Excluding period prevalence measurements reduces the quantity of data and produces results that do not reflect the regional variation present in the excluded data. But including the period prevalence measurements without a covariate to adjust for their systematic bias leads to estimates that are noticeably higher in regions where there are data on point and period prevalence. Using a fixed effect on a period prevalence indicator covariate allows the model to use all available data and explain the systematic bias and variation that result from different recall periods, as seen in figure 14.2.

The results of the model with a fixed effect on recall period show that studies on period prevalence typically measure prevalence levels that are 49% (95% UI: [12, 91]) higher than if they measured point prevalence.

A limitation of applying this method to the global data set is that it assumes the cross-walk factor is identical over age and sex for all regions of the world. This can be addressed by different cross-walks for different ages and sexes. In practice, there is rarely enough data to move beyond this assumption. However, future applications may benefit from random effects.
Figure 14.2. Comparison of prevalence estimates for anxiety disorders in 2005 in Australasian females using point prevalence data only, and using point and period prevalence data with and without a fixed effect.

or modeling interactions between cross-walk covariates and age, sex, time, or geography.
Chapter 15

Improving out-of-sample prediction: liver cirrhosis

Ali Mokdad, Abraham D. Flaxman, Hannah M. Peterson, Christopher J. L. Murray, and Mohsen Naghavi

Besides explaining the bias of noisy measurements as discussed in chapter [14] fixed effects can also increase the accuracy of out-of-sample predictions. By modeling the relationship between the epidemiological parameter of interest and an explanatory covariate, the model can extrapolate estimates for areas where no direct measurements are available, using the inferred relationship with the known covariate data. For example, only a few regions have direct measurements for the prevalence of cirrhosis of the liver. However, by using the natural log of the age-standardized cirrhosis death rate as a country-level covariate to predict out-of-sample, it is possible to estimate the prevalence of cirrhosis in regions where there has been no direct measurement. Unlike the random-effects approach in chapter [13] this approach attempts to explain the levels of the national variation in this disease and not only capture their magnitude.

Cirrhosis of the liver is the result of chronic liver damage and is characterized by an advanced stage of liver fibrosis. Cirrhosis is the end stage of any chronic liver disease, with the most common causes being alcoholic liver disease and hepatitis B and C infections. Asymptomatic until a late stage of the disease, “compensated cirrhosis” may go undetected until complications develop. The diagnostic gold standard for cirrhosis is a liver biopsy. Complications such as jaundice, ascites, esophageal varices, and liver failure mark the progression from compensated to decompensated cirrhosis. The damage is irreversible, and cirrhosis management involves the prevention,
control, and treatment of cirrhosis complications, with liver transplantation being the ultimate treatment. Without a liver transplant, mortality from decompensated cirrhosis is very high.\[105][106][107]

Hospital databases yielded prevalence and cause-specific mortality rate data from 4 (of 21 possible) regions (figure 15.1). Given the difficulty in cirrhosis diagnosis, it is assumed that these data represent decompensated cirrhosis, and the following analysis focuses on the symptomatic, decompensated phase of the disease.

**Figure 15.1.** Available global data for cirrhosis (a) prevalence and (b) cause-specific mortality.

Since decompensated cirrhosis is a very severe condition, it is a sensible hypothesis that there is a strong association between prevalence and cause-specific mortality rate at the country level. In other words, we expected a priori that regions with higher death rates from cirrhosis would also have higher decompensated cirrhosis prevalence. Figure 15.2 shows the relationship between decompensated cirrhosis prevalence and the age-standardized death rate (ASDR) of cirrhosis as a scatterplot, using all data on cirrhosis prevalence collected in systematic review.

In regions with no cirrhosis data, estimates of the ASDR may be used as an explanatory covariate to estimate cirrhosis prevalence. By borrowing strength from the mortality estimates to inform the incidence estimates, cirrhosis prevalence can be estimated for regions without data as shown in figure 15.3. Since the predictive covariate fixed effects are modeled as a
Figure 15.2. Relationship between cirrhosis prevalence data collected in systematic review and age-standardized death rate of cirrhosis.

shift in log-space, it is often preferable to use the log of the ASDR as a covariate instead of using the ASDR untransformed.

As shown in figure 15.3(b), despite the absence of any direct measurements of cirrhosis prevalence in Egypt, this approach provides a sensible estimate. It shows that the prevalence is much higher there than in the USA. It also shows a sizable uncertainty interval, reflecting the imperfect relationship between prevalence and the ASDR. The age pattern is informed by borrowing strength from regions like North America where age-specific data are available.

Even in settings where there is not such a strong explanatory relationship between the parameter of interest and some covariate, this approach can still be useful. Although it would be ideal to have direct measurements
Figure 15.3. Cirrhosis prevalence estimates and available data for (a) the USA and (b) Egypt, for males in 2005. Note that systematic review found no measurements of cirrhosis prevalence for Egypt, and prevalence level is extrapolated based on the inferred relationship with the age-standardized death rate.

of the quantities of interest, in many cases they have never been made. Explaining regional variation with a weak covariate is preferable to not explaining it at all.
Chapter 16

Risk factors: fruit consumption

Stephen S. Lim, Hannah M. Peterson, and Abraham D. Flaxman

Although the primary focus of the metaregression framework developed in this book is on estimating prevalence of disease, it is also useful for estimating other age-specific quantities of interest in descriptive epidemiology. Subsequent chapters will address the consistent estimation framework that allows data on incidence, remission, and mortality to be included in estimates of prevalence; this approach produces estimates of age-specific incidence, remission, and mortality as ancillary outputs. But before turning to that, this chapter considers the possibility of using the age-standardizing mixed-effects spline model to produce estimates of risk factor exposures.

An important component of GBD 2010 was the estimation of disease burden attributable to risk factors, such as the lack of fresh fruit. Fruit consumption is a nonnegative quantity, so it is acceptable to model it with the negative-binomial rate model. This is somewhat inelegant, however, because this model assumes that the underlying quantity being measured is count data. A count model is appropriate for prevalence and incidence rates, which represent, for example, how many cases were observed during a certain period of observation. However, a continuous distribution, such as one of the transformed normal models from chapter 2, is more common for modeling a quantity like kilograms of fruit consumed per day per capita. In this chapter, we will compare the results of using the negative-binomial rate model with two more traditional models for continuous data: the normal model and the lognormal model.

Fruit consumption has a significant protective effect against morbidity and mortality from several diseases. Measured as the total intake of fruit
per day (kg/d), fruit includes all fresh, frozen, cooked, canned, or dried
fruits, excluding fruit juices and salted or pickled fruits.\textsuperscript{108,109}

Systematic review for risk factor epidemiology proceeds much the same
as for disease prevalence, at least for determining the population-level ex-
posure to the risk. In the case of lack of fruit, systematic review collected
1502 data points on age-specific consumption of fruit.

Often it is useful to design statistical models based on real-world pro-
cesses, using count models for discrete data and rate models for continuous
data as discussed in chapter 2. Since fruit consumption is a continuous
variable, one may choose the lognormal or normal rate model over the
negative-binomial model to maintain a mechanistic understanding of the
statistical model. The models differ in their treatment of numbers that
are very close to zero, but, with well-behaved data, the model estimates
are largely independent of the choice of rate model, as seen in figure 16.1,
which compares estimates for models fitted to data from the USA in 2005.

![Figure 16.1](image)

Figure 16.1. Estimates of fruit fruit consumption in males in the USA in 2005.
(a) shows age-specific estimate using negative binomial model, (b) and (c) compare
the posterior distribution of age-standardized consumption estimate of the negative-
binomial model to the estimates using the lognormal and normal rate models.

In a setting where the data are sparser and noisier, the models still yield
quite similar results, as shown in figure 16.2 for data from Western Europe
2005. It is worth noting that unlike the estimates in figure 16.1, figure 16.2
estimates do not go through the data at the youngest ages (0–20 years). This is because fruit consumption has substantial country-level random effects, shown with a comparison of Iceland and Greece in figure 16.3.

Figure 16.2. Fruit consumption estimates in Greece and Iceland males in 2005 using the negative-binomial rate models.

In figure 16.3, small differences in the estimates at the country level are noticeable. Table 16.1 shows that the mean age-standardized country estimates differ by rate type. However, while estimates at the country level differ slightly, the regional estimate remains the same.

The country-level estimate for Greece also highlights a challenge for this approach. Since there are no data from Greece for the youngest age groups, the model borrows strength from other countries in the region. But since these countries are lower than average in adult consumption, while Greece is higher, the model predicts very high consumption in children. Assuming the
same relationship in country-to-country variation for children as for adults seems reasonable in theory, and indeed there are no data to contradict this; however, the resulting estimates are substantially above any of the levels measured in the region. This is suspicious and certainly justifies additional investigation. It could also call for incorporating additional priors into the model, based on expert knowledge about the absolute levels, country-to-country variation, or smoothness or monotonicity of the age pattern.
Table 16.1. Random-effect estimates from hierarchical spline models of fruit consumption with differing rate models

<table>
<thead>
<tr>
<th>Model</th>
<th>Iceland [95% UI]</th>
<th>Greece [95% UI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative-binomial</td>
<td>−0.57 [−0.8, −0.4]</td>
<td>0.81 [0.7, 1.0]</td>
</tr>
<tr>
<td>Lognormal</td>
<td>−0.55 [−0.7, −0.4]</td>
<td>0.80 [0.6, 1.0]</td>
</tr>
<tr>
<td>Normal</td>
<td>−0.41 [−0.6, −0.2]</td>
<td>0.78 [0.7, 0.9]</td>
</tr>
</tbody>
</table>

With ample data, different rate models produce small differences in moderately noisy data, mostly at the country level. Similar estimates from the negative-binomial, normal, and lognormal rate models only build confidence that the model is not sensitive to the choice of these rate types. When the data are sparser or noisier, considering the difference between estimates produced by different models can be a useful component of a sensitivity analysis.
Chapter 17

The compartmental model: end-stage renal disease

Sarah K. Wolf, Abraham D. Flaxman, Mohsen Naghavi, and Giuseppe Remuzzi

We now turn our attention to conditions where systematic review uncovered a substantial amount of nonprevalence data, which we wish to use to inform our estimates. This was already touched on in chapter 12 and will be investigated more systematically in the next few chapters. We begin by considering end-stage renal disease (ESRD) treated with dialysis, a condition for which data on prevalence, incidence, remission, and with-condition mortality were all collected in relatively large quantities through systematic review.

ESRD is the final stage of chronic kidney disease (CKD), the slow and progressive loss of kidney function. The most common causes of CKD are diabetes and high blood pressure. Damage to kidneys is usually permanent, but treatment and lifestyle changes can slow disease progression. However, at the final stage of the disease, the kidneys no longer function, and the patient needs dialysis or a kidney transplant to survive. There are two main types of dialysis for kidney treatment: hemodialysis and peritoneal dialysis. Hemodialysis filters waste and excess fluids from the blood externally using a machine, while peritoneal dialysis uses the lining of the peritoneal cavity and a catheter to filter wastes from the blood.

This example focuses on ESRD treated with dialysis, combining hemodialysis and peritoneal dialysis for analysis. Most of the data are from studies or registry reports. Transplantation incidence among the prevalent dialysis population was used as a proxy for remission. The analysis includes 5664
As discussed in section 7.2, epidemiological parameters, such as incidence, prevalence, remission, and with-condition mortality, are related by a logical requirement of internal consistency. A prevalent case can exist only if there was a past incident event, and the current number of prevalent cases can be determined from past prevalent cases, newly incident cases, deaths, and remissions. Modeling the parameters simultaneously produces a best
estimate and plausible uncertainty bounds for incidence and prevalence that are internally consistent for a single time, place, and sex.

Figure 17.2 compares the compartmental and spline model results for Australasian males with ESRD treated with dialysis in 2005. While the spline model estimates each epidemiological parameter individually, the compartmental model estimates prevalence, incidence, remission, and withdrawal mortality simultaneously. Figure 17.2 and a comparison of the age-standardized prevalence estimates for the region show that the compartmental model estimates do not follow the data like the spline model does. As seen in figure 17.3, the spline model produces prevalence estimates that are systematically lower than those of the compartmental model because of the logic requirement that requires all prevalent cases to have a corresponding incident event.

Another advantage to compartmental modeling is an estimate with a smooth age pattern. Modeling each epidemiological parameter individually, the spline model follows the data exactly, often producing an uneven age pattern as seen in figure 17.4. This effect can be minimized by placing an informative prior on the penalized spline model as discussed in chapter 3.

The compartmental model is preferable to modeling each parameter individually with the spline model because it incorporates all available data. Simultaneously modeling all data, the compartmental model produces internally consistent estimates for a single age, sex, and time.
Figure 17.2. Comparison of epidemiological parameter estimates for Australasian males with ESRD treated with dialysis in 2005 using the compartmental and spline models.
Figure 17.3. Comparison of the regional age-standardized prevalence estimates using compartmental and spline models for males with ESRD treated with dialysis in 2005.
Figure 17.4. With-condition mortality estimates for Australasian males with ESRD treated with dialysis in 2005 using a compartmental model, a spline model, and a spline model with a smoothing parameter.
Chapter 18

Knot selection in compartmental spline models: osteoarthritis of the knee

Marita Cross, Damian Hoy, Theo Vos, Abraham D. Flaxman, and Lyn March

Chapters 9 and 10 demonstrated the importance of knot selection in spline models for sparse, noisy data. In this chapter, we return to this point in the context of compartmental models where the age-specific hazards are represented by splines and the parameter of primary interest, prevalence, comes from the solution to a system of differential equations based on these splines. In this setting, modeling decisions about the knot locations for one parameter affect the estimates for all the other parameters as well. The compartmental model for osteoarthritis of the knee provides a demonstration of this.

Osteoarthritis (OA) is a joint disorder that affects joint cartilage and underlying bone. OA causes pain in the joints and limits movement. OA of the knee is common and causes significant morbidity, particularly in the elderly. Systematic review yielded 602 data points representing 27 countries in 10 GBD 2010 regions.

Since OA of the knee is exceedingly rare in young adults, expert priors inform the model that the onset of the disease does not start before age 30. The number and location of knots in the incidence rate after this minimum age of onset determine critical features of the model. As shown in figure 18.1, it is important to have enough knots to represent the rapid
change in age-specific incidence. However, knot selection in incidence also affects estimates of prevalence and excess mortality.

**Figure 18.1.** Knot selection between the ages of 30 and 99 plays an important role in the estimates of (a) prevalence, (b) incidence, and (c) excess mortality for Western European females with OA of the knee in 2005. The incidence rate of all models has knots at \{0, 30, 40, 45, 65, 100\}. Between the ages of 30 and 40, the models have either no additional knots (\{\}) or additional knots at \{35\} or \{31, 35\}.

The model is also sensitive to assumptions about the epidemiological profile, expressed in the model as expert priors. Figure 18.2 compares assumptions about OA of the knee incidence. A prior that requires zero incidence at ages greater than 99 implies that incidence decreases with age. In other words, after a certain age, if OA of the knee has not developed,
it is unlikely it ever will. Without this prior, incidence increases with age. The logic requirement of internal consistency in the compartmental model means that prevalence estimates are also affected, as shown in figure 18.2. When incidence is unrestricted, prevalence has a very different age pattern than when incidence is restricted.

![Figure 18.2. A comparison of (a) prevalence and (b) incidence estimates for Western European females with OA of the knee in 2005 with and without a prior stipulating no onset of the disease in ages greater than 99 using a compartmental model.](image)

Knot selection in compartmental models can be an influential part of the model. The logic requirement of internal consistency means that modeling decisions for one parameter affect all parameters. Knot selection had a substantial effect on the estimates, although it can be minimized by increasing the number of knots included in the model. Likewise, informative priors on incidence drastically modified other parameter estimates. With ample data, the consistent model is robust to knot selection. When there is not enough data, knot selection should be based on principled considerations and a sensitivity analysis performed. The sensitivity analysis shows the effect of assumptions and the range of results, allowing the modeler to understand the differences arising from different assumptions. The influence of priors in compartmental modeling is elaborated further in the following chapter.
Chapter 19

Expert priors in compartmental models: bipolar disorder

Alizé Ferrari, Abraham D. Flaxman, Hannah M. Peterson, Theo Vos, and Harvey Whiteford

Just as priors in the spline model were influential on the estimates produced for PMS prevalence in chapter 10, the priors on the age-specific rates in a compartmental model can be influential on the estimates. The situation is more complicated here, however, because priors on a hazard of one type propagate through to affect the estimates for all other parameters as well, due to the consistency enforced by the compartmental model. In this chapter, we will use the meta-analysis of bipolar disorder prevalence as an example of the effects of informative priors on levels of age-specific incidence and remission hazards.

Bipolar disorder is a mental disorder that causes the experience of at least one manic and one major depressive episode, interspersed by periods of residual symptoms. A manic episode is characterized by elevated, expansive, or irritable mood, while a depressive episode is characterized by depressed mood or loss of interest in everyday activities. A shift between episodes is demarcated by either a change in symptoms to the opposite polarity or experiences of residual symptoms that are below the threshold for a manic or depressive episode. In the case of rapid cycling, shifts between episodes occur as frequently as four or more shifts in a given year. Extreme behavior changes accompany mood changes, and it is not uncommon for sleeping, eating, or activity patterns to change with manic and depressive
episodes. While there is no cure, treatment helps manage mood swings and related symptoms.\cite{114,115}

The modeling of bipolar disorder is based on literature describing it as a chronic illness with little or no complete remission. The terms “residual” and “remission” have very different implications for GBD 2010. A residual state involves less severe symptoms with lesser disability but still contributes to disease burden. Remission is equivalent to a cure rather than a temporary reduction in symptom levels and thus does not contribute to burden. No studies were found reporting on complete remission from bipolar disorder as per this definition, which is consistent with the description in the literature that there is no cure.\cite{86} In this chapter, analysis uses only data from high-income North America shown in figure \ref{fig:19.1}.

While there is evidence to suggest that bipolar disorder commonly starts in the midteens or early twenties, there is still disagreement over a minimum age of onset. Even though symptoms can be tracked back to childhood, setting a threshold for diagnosis is difficult given that current diagnostic criteria are based on the adult presentation of the disorder. Literature and expert advice suggest that although prepubertal bipolar disorder is rare, there is a possibility it may exist.\cite{114,115}

While expert priors are useful in guiding the modeling process, they may have unintended effects, as discussed in chapter \ref{chapter:4}. Choosing to have no restrictions on the age of onset alters the age-specific prevalence greatly, as shown in figure \ref{fig:19.2}.

Like the age of onset, little is known about the upper age limit of bipolar disorder. Using expert knowledge to set plausible bounds on the level of disease is useful in modeling noisy data. However, changes in the upper age limit may produce unexpected changes, as shown in figure \ref{fig:19.3}. The prevalence estimates in figure \ref{fig:19.3} are about the same because there are enough data to inform the model, but incidence, remission, and excess mortality undergo subtle changes to account for the prior.

In sparse and noisy data, sometimes the changes to account for the prior are not so subtle, as shown in the sensitivity analysis in figure \ref{fig:19.4}. Here, small changes in the prior level on remission lead to large changes in the estimated excess mortality.

The internal consistency in the compartmental model causes modeling decisions, such as priors on level, for one parameter to propagate and affect all other parameter estimates. When working with ample data, the model estimates are robust against the choice of informative priors on level. How-
Figure 19.1. Data on (a) prevalence, (b) incidence, (c) remission, and (d) standardized mortality ratio collected from systematic review of bipolar disorder in high-income North America.

ever, these choices can cause substantial changes to estimates when working with sparse and noisy data.
Figure 19.2. Estimates of the prevalence of bipolar disorder for males in the GBD 2010 region of high-income North America in 1990 using differing priors for age of onset in a compartmental model.
Figure 19.3. Estimated (a) prevalence, (b) incidence, (c) remission, and (d) excess mortality for males with bipolar disorder in the GBD 2010 region of high-income North America in 1990 using a compartmental model with different priors on the upper age limit of incidence of 45 years, 65 years, or unrestricted.
Figure 19.4. Estimated (a) remission and (b) excess mortality for bipolar disorder in males in high-income North America in 1990 in a compartmental model with different priors on remission that limit remission to 0, 5, or 10 per 100 PY.
Chapter 20

Cause-specific mortality rates: alcohol dependence

Theo Vos, Jed Blore, Abraham D. Flaxman, Hannah M. Petersen, and Juergen Rehm

A key assumption of the YLL estimates in GBD 2010 is that only one cause leads to death. This categorical attribution of deaths in a mutually exclusive way has the desirable property that all assigned deaths sum to the total number of deaths. However, this creates a rate that does not directly correspond to any rate in the compartmental model (figure 7.3). Using this cause-specific mortality rate (CSMR) as a measurement of \( h_p \cdot h_f \) forces the assumption that all who die with the condition die of the condition. While this is a reasonable assumption for conditions such as cancer, cirrhosis, or diarrhea, it is not for conditions such as alcohol dependence, where many people die with the condition but not of it. Using alcohol dependence as an example, this chapter compares the results of the modeling assumptions of those who die of alcohol dependence and those who die with alcohol dependence but of other causes.

Alcohol dependence is the dysfunctional pattern of alcohol consumption that leads to physiological dependence and impaired control. Similar to cocaine dependence (chapter 9), in order to be diagnosed with alcohol dependence, 3 or more of the 7 substance dependence criteria identified by the American Psychiatric Association must be fulfilled within a 12-month period. Systematic review yielded prevalence, excess mortality, and cause-specific mortality data as seen in figure 20.1.

To include cause-specific mortality data in the compartmental model, the model in figure 7.3 can be adapted by splitting the excess mortality \( h_f \)
into two parts (figure 20.2): those who die of the condition $h_f'$ and those who die with the condition but not of it $h_f''$. As described in section 2.7, excess mortality can then be represented as

$$h_f = h_f'' + h_f'$$

While the product of excess mortality and prevalence, $h_p \cdot h_f$, can be directly measured, in practice $h_f''$ and $h_f'$ are never disentangled. Our method implicitly separates $h_f'$ and $h_f''$ but does not try to explicitly represent both.
For some diseases in GBD 2010, it is a reasonable assumption that $h_{f''} = 0$, so that those who die with the condition die of it. When this is the case, the cause-specific mortality is a direct estimate of $h_p \cdot h_f$, a measurement of population mortality. However, for diseases such as alcohol dependence, this is a questionable assumption, for the reasons mentioned above. When $h_{f''} \neq 0$, cause-specific mortality data provide the lower bound on $h_p \cdot h_f$.

As seen in figure 20.3, assuming $h_{f''} = 0$ instead of $h_{f''} \geq 0$ leads to a change in the estimated prevalence of the condition.

By decomposing excess mortality into those who die of the disease and those who die with the disease but not of the disease, the compartmental model in figure 7.3 can use the cause-specific mortality data as a lower bound on prevalence times excess mortality, which is appropriate if the disease is not exclusively coded as the underlying cause of death.
Figure 20.3. Comparison of (a) prevalence, (b) cause-specific mortality, and (c) excess mortality estimates of alcohol dependence in Central Asian males in 2005 when using cause-specific mortality data as a lower bound ($h_f''$ unrestricted) or a direct estimate of the product of prevalence and excess mortality ($h_f'' = 0$).
Chapter 21

Conclusion

Abraham D. Flaxman, Christopher J. L. Murray, and Theo Vos

This book has developed our new metaregression framework for descriptive epidemiology from first principles, and applied it to several examples. This approach is new and necessarily complicated. However, in exposition we strove to make it no more complicated than necessary.

There are 7 ways that this approach differs from traditional metaregression methods, each addressing commonly occurring features of the data collected in systematic review during GBD 2010. Recall that the data are often very sparse and very noisy. When there are whole regions of the globe for which no data are available; when there are not measurements of prevalence, but only of incidence; when the regions that do have data have measurements that vary 10 times more than sampling error would allow; and in many combinations of these challenging environments, we must produce estimates that reflect the uncertainty of the available data.

Our approach has used:

- The negative binomial model of data, an approach that allows groups to have observations of zero and dispersion beyond that seen in Poisson and binomial models.

- A piecewise linear spline model of age-specific hazards, which balances computational tractability and age-pattern flexibility.

- Bayesian methods, to quantify uncertainty and incorporate priors based on biology, exposure, or clinical series.

- The age-standardizing model of heterogeneous, nonstandard age groups, such as 18–35 or 15 and older.
CHAPTER 21. CONCLUSION

- Fixed effects modeling to crosswalk between available studies that use different case definitions, to predict out-of-sample, based on known country-specific covariates, and to quantify differences in nonsampling variance in different study types.

- Random effects modeling to capture true variation within and between regions.

- Integrative systems modeling to combine data collected for many different outcomes, such as incidence, prevalence, remission, excess mortality, or cause-specific mortality.

There are a number of important extensions to this approach and other avenues for future work that should be pursued in the future. These have been highlighted at the end of each chapter in Part I of the book, and are all collected here, as a plan for how to move forward.

Now that systematic reviews have been conducted for numerous conditions, the data gathered in the reviews can be used for a systematic out-of-sample cross-validation exercise to refine modeling choices and suggest new improvements.

Further research into the rate models from chapter 2 can continue to develop and test the offset log-transformed model and see how it compares to the negative binomial model on a range of models collected through systematic review. Perhaps with additional innovation in computation algorithms, the beta-binomial model can also be made efficient enough to be seriously considered as well, but this seems to require methods beyond the MCMC approach of chapter 8.

The age-specific hazard functions in chapter 3 have more model parameters than ideal. Future work should be dedicated to removing the necessity of spline-modeling decisions about the number of knots, location of knots, and level of smoothing. If computational algorithms become fast enough, this could be accomplished simply by exploring a range of parameters and selecting (or averaging) based on out-of-sample predictive validity. An alternative line of research would be to go from spline models to Gaussian processes, or some related nonparametric formulation of age-specific hazard function. This will certainly have its own computational challenges.

The priors in chapter 4 can be extended through three directions of future work. First, the expert priors deserve an automated and fool-proof procedure for sensitivity analysis, so the impact of these assumptions can
be assessed quickly and comprehensively. Second, there are additional for-
formulations of expert priors which would be helpful in certain settings, such
as the unimodality prior described at the end of that chapter. Third, and
most importantly, the empirical priors that have been used so far must
be compared to alternative formulations, particularly in terms of the as-
sumptions about the variance-covariance matrix for the joint distribution
of hazards at different ages.

Age group modeling and covariate modeling can be extended to consider
alternative models of heterogeneous age groups, uncertainty, and missing-
ness in predictive covariates. Based on the success of ensemble modeling for
covariates in other settings, that approach could also be considered here,
but only if the computation time for individual models in the ensemble can
be sped up substantially first.

Removing the endemic equilibrium assumption from section 7.3 is an-
other important direction for future work. The available data for most
global descriptive epidemiology is too sparse and noisy to justify this compli-
cation, but for many national and subnational analyses, this will be highly
relevant.

All of these directions for future work will be facilitated by research
into numerical algorithms and computational infrastructure, since the time
it takes to run these models is currently an impediment to large cross-
validation exercises as well as searching for optimal parameter settings
through out-of-sample predictive validity.

If substantial speed increases are possible through improved algorithms
and infrastructure, it may also be possible to go beyond the empirical Bayes
approach and to fit a full hierarchical model for all regions (or countries)
simultaneously. This will require innovation in models as well as methods to
determine the most appropriate way to formalize the hierarchical similarity
priors. It is likely to be extremely demanding computationally, however,
and it could be that experimenting with alternative approaches to the em-
pirical prior model will yield an intermediate approach that requires only a
more modest increase in computational power.

Besides computational algorithms and speed, another area of research is
to create more tools for the analyst to compare different rate types and the
effects of modeling decisions. In particular, developing a way to determine
the effect of the priors and the effect of the data on the posterior estimate
would be valuable for model checking.

The development of a metaregression approach for descriptive epidemi-
ological prediction has been quite an adventure. It is likely to continue to be in the future as data, models, methods, and infrastructure change and improve, making even more precise and accurate estimation possible.
Appendix A

GBD Study 2010 spatial hierarchy
<table>
<thead>
<tr>
<th>GBD superregion</th>
<th>GDB region</th>
<th>Country</th>
<th>ISO3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Armenia</td>
<td>ARM</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Azerbaijan</td>
<td>AZE</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Georgia</td>
<td>GEO</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Kazakhstan</td>
<td>KAZ</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Kyrgyzstan</td>
<td>KGZ</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Mongolia</td>
<td>MNG</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Tajikistan</td>
<td>TJK</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Turkmenistan</td>
<td>TKM</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Uzbekistan</td>
<td>UZB</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Albania</td>
<td>ALB</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Asia</td>
<td>Bosnia and Herzegovina</td>
<td>BIH</td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>-----------------------------------------------------</td>
<td>----------------</td>
<td>------------------</td>
<td>------</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Bulgaria</td>
<td>BGR</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Croatia</td>
<td>HRV</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Czech Republic</td>
<td>CZE</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Hungary</td>
<td>HUN</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Macedonia</td>
<td>MKD</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Montenegro</td>
<td>MNE</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Poland</td>
<td>POL</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Romania</td>
<td>ROU</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Serbia</td>
<td>SRB</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Slovakia</td>
<td>SVK</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Central Europe</td>
<td>Slovenia</td>
<td>SVN</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Eastern Europe</td>
<td>Belarus</td>
<td>BLR</td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>---------------------------------------------------------------------</td>
<td>-------------------</td>
<td>-----------</td>
<td>------</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Eastern Europe</td>
<td>Estonia</td>
<td>EST</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Eastern Europe</td>
<td>Latvia</td>
<td>LVA</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Eastern Europe</td>
<td>Lithuania</td>
<td>LTU</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Eastern Europe</td>
<td>Moldova</td>
<td>MDA</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Eastern Europe</td>
<td>Russia</td>
<td>RUS</td>
</tr>
<tr>
<td>Central Europe, Eastern Europe, and Central Asia</td>
<td>Eastern Europe</td>
<td>Ukraine</td>
<td>UKR</td>
</tr>
<tr>
<td>High-income</td>
<td>Australasia</td>
<td>Australia</td>
<td>AUS</td>
</tr>
<tr>
<td>High-income</td>
<td>Australasia</td>
<td>New Zealand</td>
<td>NZL</td>
</tr>
<tr>
<td>High-income</td>
<td>High-income Asia Pacific</td>
<td>Brunei</td>
<td>BRN</td>
</tr>
<tr>
<td>High-income</td>
<td>High-income Asia Pacific</td>
<td>Japan</td>
<td>JPN</td>
</tr>
<tr>
<td>High-income</td>
<td>High-income Asia Pacific</td>
<td>South Korea</td>
<td>KOR</td>
</tr>
<tr>
<td>High-income</td>
<td>High-income Asia Pacific</td>
<td>Singapore</td>
<td>SGP</td>
</tr>
<tr>
<td>High-income</td>
<td>High-income North America</td>
<td>Canada</td>
<td>CAN</td>
</tr>
<tr>
<td>High-income</td>
<td>High-income North America</td>
<td>USA</td>
<td>USA</td>
</tr>
<tr>
<td>High-income</td>
<td>Southern Latin America</td>
<td>Argentina</td>
<td>ARG</td>
</tr>
<tr>
<td>High-income</td>
<td>Southern Latin America</td>
<td>Chile</td>
<td>CHL</td>
</tr>
<tr>
<td>High-income</td>
<td>Southern Latin America</td>
<td>Uruguay</td>
<td>URY</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Andorra</td>
<td>AND</td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>------------------------</td>
<td>----------------</td>
<td>------</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Austria</td>
<td>AUT</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Belgium</td>
<td>BEL</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Cyprus</td>
<td>CYP</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Denmark</td>
<td>DNK</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Finland</td>
<td>FIN</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>France</td>
<td>FRA</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Germany</td>
<td>DEU</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Greece</td>
<td>GRC</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Iceland</td>
<td>ISL</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Ireland</td>
<td>IRL</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Israel</td>
<td>ISR</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Italy</td>
<td>ITA</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Luxembourg</td>
<td>LUX</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Malta</td>
<td>MLT</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Netherlands</td>
<td>NLD</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Norway</td>
<td>NOR</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Portugal</td>
<td>PRT</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Spain</td>
<td>ESP</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Sweden</td>
<td>SWE</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>Switzerland</td>
<td>CHE</td>
</tr>
<tr>
<td>High-income</td>
<td>Western Europe</td>
<td>UK</td>
<td>GBR</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Andean Latin America</td>
<td>Bolivia</td>
<td>BOL</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Andean Latin America</td>
<td>Ecuador</td>
<td>ECU</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Andean Latin America</td>
<td>Peru</td>
<td>PER</td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------</td>
<td>--------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Antigua and Barbuda</td>
<td>ATG</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Bahamas</td>
<td>BHS</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Barbados</td>
<td>BRB</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Belize</td>
<td>BLZ</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Bermuda</td>
<td>BMU</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Cuba</td>
<td>CUB</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Dominica</td>
<td>DMA</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Dominican Republic</td>
<td>DOM</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Grenada</td>
<td>GRD</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Guyana</td>
<td>GUY</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Haiti</td>
<td>HTI</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Jamaica</td>
<td>JAM</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Puerto Rico</td>
<td>PRI</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Saint Lucia</td>
<td>LCA</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Saint Vincent and the</td>
<td>VCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Grenadines</td>
<td></td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Suriname</td>
<td>SUR</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Caribbean</td>
<td>Trinidad and Tobago</td>
<td>TTO</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Colombia</td>
<td>COL</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Costa Rica</td>
<td>CRI</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>El Salvador</td>
<td>SLV</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Guatemala</td>
<td>GTM</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Honduras</td>
<td>HND</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Mexico</td>
<td>MEX</td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------------------</td>
<td>-------------------</td>
<td>------</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Nicaragua</td>
<td>NIC</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Panama</td>
<td>PAN</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Central Latin America</td>
<td>Venezuela</td>
<td>VEN</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Tropical Latin America</td>
<td>Brazil</td>
<td>BRA</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>Tropical Latin America</td>
<td>Paraguay</td>
<td>PRY</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Afghanistan</td>
<td>AFG</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Algeria</td>
<td>DZA</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Bahrain</td>
<td>BHR</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Egypt</td>
<td>EGY</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Iran</td>
<td>IRQ</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Iraq</td>
<td>JOR</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Jordan</td>
<td>KWT</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Kuwait</td>
<td>LBN</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Lebanon</td>
<td>LBY</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Libya</td>
<td>MAR</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Morocco</td>
<td>PSE</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Occupied Palestinian</td>
<td>OMN</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Oman</td>
<td>QAT</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Qatar</td>
<td>SAU</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Saudi Arabia</td>
<td>SYR</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Syria</td>
<td>TUN</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Tunisia</td>
<td>TUR</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Turkey</td>
<td></td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>United Arab Emirates</td>
<td>ARE</td>
</tr>
<tr>
<td>North Africa and Middle East</td>
<td>North Africa and Middle East</td>
<td>Yemen</td>
<td>YEM</td>
</tr>
<tr>
<td>South Asia</td>
<td>South Asia</td>
<td>Bangladesh</td>
<td>BGD</td>
</tr>
<tr>
<td>South Asia</td>
<td>South Asia</td>
<td>Bhutan</td>
<td>BTN</td>
</tr>
<tr>
<td>South Asia</td>
<td>South Asia</td>
<td>India</td>
<td>IND</td>
</tr>
<tr>
<td>South Asia</td>
<td>South Asia</td>
<td>Nepal</td>
<td>NPL</td>
</tr>
<tr>
<td>South Asia</td>
<td>South Asia</td>
<td>Pakistan</td>
<td>PAK</td>
</tr>
<tr>
<td>South Asia</td>
<td>East Asia</td>
<td>China</td>
<td>CHN</td>
</tr>
<tr>
<td>South Asia, East Asia, and Oceania</td>
<td>East Asia</td>
<td>Hong Kong Special Administrative Region of China</td>
<td>HKG</td>
</tr>
<tr>
<td>South Asia, East Asia, and Oceania</td>
<td>East Asia</td>
<td>North Korea</td>
<td>PRK</td>
</tr>
<tr>
<td>South Asia, East Asia, and Oceania</td>
<td>East Asia</td>
<td>Macao Special Administrative Region of China</td>
<td>MAC</td>
</tr>
<tr>
<td>South Asia, East Asia, and Oceania</td>
<td>East Asia</td>
<td>Taiwan</td>
<td>TWN</td>
</tr>
<tr>
<td>South Asia, East Asia, and Oceania</td>
<td>Oceania</td>
<td>Fiji</td>
<td>FJI</td>
</tr>
<tr>
<td>South Asia, East Asia, and Oceania</td>
<td>Oceania</td>
<td>Kiribati</td>
<td>KIR</td>
</tr>
<tr>
<td>South Asia, East Asia, and Oceania</td>
<td>Oceania</td>
<td>Marshall Islands</td>
<td>MHL</td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------------</td>
<td>------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and</td>
<td>Oceania</td>
<td>Federated States of Micronesia</td>
<td>FSM</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Papua New Guinea</td>
<td>PNG</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Samoa</td>
<td>WSM</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Solomon Islands</td>
<td>SLB</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Tonga</td>
<td>TON</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Vanuatu</td>
<td>VUT</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and</td>
<td>Southeast Asia</td>
<td>Cambodia</td>
<td>KHM</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Indonesia</td>
<td>IDN</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and</td>
<td>Southeast Asia</td>
<td>Laos</td>
<td>LAO</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Malaysia</td>
<td>MYS</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and</td>
<td>Southeast Asia</td>
<td>Maldives</td>
<td>MDV</td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td>Mauritius</td>
<td>MUS</td>
</tr>
</tbody>
</table>
Table A.1. (continued)

<table>
<thead>
<tr>
<th>GBD superregion</th>
<th>GDB region</th>
<th>Country</th>
<th>ISO3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southeast Asia, East Asia, and Oceania</td>
<td>Southeast Asia</td>
<td>Burma</td>
<td>MMR</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and Oceania</td>
<td>Southeast Asia</td>
<td>Philippines</td>
<td>PHL</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and Oceania</td>
<td>Southeast Asia</td>
<td>Seychelles</td>
<td>SYC</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and Oceania</td>
<td>Southeast Asia</td>
<td>Sri Lanka</td>
<td>LKA</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and Oceania</td>
<td>Southeast Asia</td>
<td>Thailand</td>
<td>THA</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and Oceania</td>
<td>Southeast Asia</td>
<td>Timor-Leste</td>
<td>TLS</td>
</tr>
<tr>
<td>Southeast Asia, East Asia, and Oceania</td>
<td>Southeast Asia</td>
<td>Vietnam</td>
<td>VNM</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Central sub-Saharan Africa</td>
<td>Angola</td>
<td>AGO</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Central sub-Saharan Africa</td>
<td>Central African Republic</td>
<td>CAF</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Central sub-Saharan Africa</td>
<td>Congo</td>
<td>COG</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Central sub-Saharan Africa</td>
<td>Democratic Republic of Congo</td>
<td>COD</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Central sub-Saharan Africa</td>
<td>Equatorial Guinea</td>
<td>GNQ</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Central sub-Saharan Africa</td>
<td>Gabon</td>
<td>GAB</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Burundi</td>
<td>BDI</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Comoros</td>
<td>COM</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Djibouti</td>
<td>DJI</td>
</tr>
<tr>
<td>GBD superregion</td>
<td>GDB region</td>
<td>Country</td>
<td>ISO3</td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------------------------</td>
<td>---------------</td>
<td>------</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Eritrea</td>
<td>ERI</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Ethiopia</td>
<td>ETH</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Kenya</td>
<td>KEN</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Madagascar</td>
<td>MDG</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Malawi</td>
<td>MWI</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Mozambique</td>
<td>MOZ</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Rwanda</td>
<td>RWA</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Somalia</td>
<td>SOM</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Sudan</td>
<td>SDN</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Tanzania</td>
<td>TZA</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Uganda</td>
<td>UGA</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Eastern sub-Saharan Africa</td>
<td>Zambia</td>
<td>ZMB</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Southern sub-Saharan Africa</td>
<td>Botswana</td>
<td>BWA</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Southern sub-Saharan Africa</td>
<td>Lesotho</td>
<td>LSO</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Southern sub-Saharan Africa</td>
<td>Namibia</td>
<td>NAM</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Southern sub-Saharan Africa</td>
<td>South Africa</td>
<td>ZAF</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Southern sub-Saharan Africa</td>
<td>Swaziland</td>
<td>SWZ</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Southern sub-Saharan Africa</td>
<td>Zimbabwe</td>
<td>ZWE</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Benin</td>
<td>BEN</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Burkina Faso</td>
<td>BFA</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Cameroon</td>
<td>CMR</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Cape Verde</td>
<td>CPV</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Chad</td>
<td>TCD</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Côte d’Ivoire</td>
<td>CIV</td>
</tr>
</tbody>
</table>
Table A.1. (continued)

<table>
<thead>
<tr>
<th>GBD superregion</th>
<th>GDB region</th>
<th>Country</th>
<th>ISO3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Gambia</td>
<td>GMB</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Ghana</td>
<td>GHA</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Guinea</td>
<td>GIN</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Guinea-Bissau</td>
<td>GNB</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Liberia</td>
<td>LBR</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Mali</td>
<td>MLI</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Mauritania</td>
<td>MRT</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Niger</td>
<td>NER</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Nigeria</td>
<td>NGA</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>SãoTomé and Príncipe</td>
<td>STP</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Senegal</td>
<td>SEN</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Sierra Leone</td>
<td>SLE</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>Western sub-Saharan Africa</td>
<td>Togo</td>
<td>TGO</td>
</tr>
</tbody>
</table>
References


[34] Manski C, Tamer E. Inference on regressions with interval data on a regressor or outcome. Econometrica. 2002;70(2):519546.


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


