University of California

Los Angeles

Learning clinical outcomes from massive observational data

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biomathematics

by

Trevor Raymond Shaddox

2016

Abstract of the Dissertation

Learning clinical outcomes from massive observational data

by

Trevor Raymond Shaddox

Doctor of Philosophy in Biomathematics

University of California, Los Angeles, 2016

Professor Marc Adam Suchard, Chair

Emerging national patient claims and electronic health record databases offer a rich frontier for learning about treatment effectiveness and clinical decision making. However, these resources present statistical and computational challenges commensurate with their promise, requiring innovative approaches for practically and efficiently extracting meaningful results. In this dissertation, I seek to address some of these challenges. First, I present a hierarchical model for learning about the relationship between treatments and multiple related adverse outcomes simultaneously, showing that this approach can reduce bias in relative risk estimates. Second, I develop a novel minorization-maximization (MM) algorithm for uncoupling the sequential Newton steps that arise within the state of the art model fitting procedure for the conditional models popular for observational studies, allowing faster, parallelized model fitting. Third, I develop a birth-death model for treatment trajectories among patients with diabetes mellitus type II. In these sections, I discuss applications to observational healthcare datasets, demonstrating how these methods work at scale.

The dissertation of Trevor Raymond Shaddox is approved.

Eli Ipp

Kenneth L. Lange

Douglas Bell

Marc Adam Suchard, Committee Chair

University of California, Los Angeles
2016

To my family

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I am very grateful to my advisor, Marc Suchard, for taking on a medical student for a summer research project and fostering my latent love of math. You convinced me that I could turn my interests into a full-time pursuit. You made sure that I never lacked for funding or interesting research projects. But, I want to especially acknowledge your ability to shed light on what makes a research project hard, interesting, and important. Thank you for guiding me and supporting me. I also thank the other members of my dissertation committee, Kenneth Lange, Douglas Bell, and Eli Ipp, for their many valuable questions and perspectives. Dr. Lange's optimization class was my first in the PhD program, and the ideas from that class have permeated my Biomathematics experience. As someone with feet firmly planted in both the clinical and biomedical informatics worlds, Dr. Bell has helped me build the translational relevance into this dissertation. Collaborating with Dr. Ipp has been very fulfilling; he has helped me see and emphasize the parts of my research that translate most directly to making a difference in patient care.

As a student who was homeschooled from kindergarten through high school, my research is a reflection of the love of learning my parents gave me. They always encouraged me to keep my mind open about pursuing a PhD, even when I thought that pathway was closed to me. Their unending, unconditional, unhesitating support has made me the better parts of who I am. Thank you. I thank Sarah for always offering a kind ear to hear my ramblings about my day and a gentle spirit to support me when the road was tougher; you helped me find balance when I had lost it.

In Block 3 of the first year of medical school, I remember researching my newly assigned my small group instructor, Carol Newton. Little did I know what an impact she would have on my career. She casually mentioned to me that medical students could join the MD/PhD program, at least once a week for 3 months. She saw how much I missed math in medical school, and she found a way for me to find it again. I am deeply saddened that she could not see that I followed the path to which she led me.

There are many others whose friendship, and wisdom has guided my path. The Biomathematics department faculty has made sure that my training was rigorous and expansive, both inside and outside the classroom. I am very grateful to Janet Sinsheimer for her unwavering support as department chair as well as for insightful perspective on both research problems and research life, to Elliot Landaw for always asking questions in the intersection of math and medicine, and to Van Savage and Tom Chou for teaching me how to think like a physicist. Thank you, David Tomita, for going above and beyond in making life as a Biomath student easier, for always advocating on my behalf in administrative matters, and for generally making Biomath a better place with your kindness and humor.

The Suchard lab graduate students and post-docs were a constant source of support and insight during my time in the lab. Without Mandev Gill, Gabriela Cybis, Max Tolkoff, Guy Baele, Lam Ho, and Yuxi Tian, I doubt I would have survived this process. In particular, Mandev Gill's ability to make me laugh after disappointing setbacks may have been the difference between my success and failure; with empathy and humor, he was first-aid for some of the more challenging moments I have experienced. I am grateful to Gabriela Cybis for helping me navigate the Biomathematics PhD process. Max was a committed sparring partner for discussing ideas. Guy always had wisdom and insight to share, about research, life, and tennis. It has been a privilege and pleasure to work with Lam and Yuxi, and the three of us did work together that was truly collaborative. I look forward to continuing that work for many years to come. But hopefully in the form of more than one paper. Not one paper for many years. No one wants that.

More broadly, I will greatly miss Kevin Keys stopping by our office and going off on a tangent about optimization. Daniel Conn's deep thinking and excellent coffee choices made our lab space a better one. Finally, Douglas Morrison's decision to come to UCLA was spectacularly awesome. His working at the desk next to mine was a lifesaver, and his presence has immeasurably improved my graduate school experience.

Some parts of this dissertation have appeared elsewhere or are under review. Chapter 3 is joint work with Patrick B. Ryan, Martijn J. Schuemie,

David Madigan, and Marc A. Suchard and has been accepted with revisions for publication to *Statistical Analysis and Data Mining*. Chapter 4 is joint work with Patrick B. Ryan, Martijn J. Schuemie, Kenneth L. Lange, and Marc A. Suchard. Chapter 5 is a collaboration with Eli Ipp, Yang Lu, Yuxi Tian, and Lam Ho.

Throughout my graduate career at UCLA, my studies and research work have been funded by three sources. First, the Systems and Integrative Biology Training Grant (NIH T32GM008185) paid fees and a generous stipend during my first year. The Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases provided my funding for years 2-3. Finally, the Medical Scientist Training Program funded me during my final year. In addition, Marc Suchard supplemented my income with graduate student researcher funds.

Vita

2010–2018	M.D., Medical Scientist Training Program David Geffen School of Medicine at UCLA.
2010	M.S., (Biomedical Informatics) Stanford University School of Medicine.
2005-2009	B.S., (Mathematical and Computational Science, with University Distinction) Stanford University.
2014, 2015	Carol Newton Biomathematics Travel Award
2014	Junior Travel Award, International Society for Bayesian Analysis
2013–Present	Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases Training Grant
2012–2013	NIH Systems and Integrative Biology Training Grant
2005	National Merit Scholar

Publications and Presentations

2015, Shaddox TR, Ryan PB, Schuemie MJ, Madigan D, and Suchard MA. Hierarchical Models for Multiple, Rare Outcomes using Massive Observational Healthcare Databases. *Statistical Analysis and Data Mining*, accepted with revisions.

2016, Shaddox TR, Lu Y, Ho LST, Tian Y, Ipp E, and Suchard MA. Introducing Moirai: A Birth-Death Model for Diabetes Treatment Escalation. Presented at the Joint Research Symposium, Los Angeles, California

2015, Shaddox TR and Suchard MA. Monstrous MCMC: Fully Bayesian Inference in Cyclops for Massive Observational Datasets. Topic-contributed talk presented at the Joint Statistical Meetings, Seattle, Washington

2015, Shaddox TR and Suchard MA. MM Optimization in Massive Observational Analysis. Presented at Bayesian Inference in Stochastic Processes, Istanbul, Turkey

2014, Shaddox TR, Lange K, Madigan D, and Suchard MA. MM Optimization in Massive Observational Analysis. Presented at the Joint Statistical Meetings, Boston, Massachusetts

2014, Shaddox TR and Suchard MA. Extending Bayesian inference in pharmacovigilance beyond point estimates with massive parallelization. Presented at the International Society for Bayesian Analysis World Meeting, Cancun, Mexico

2013, Shaddox TR and Suchard MA. It runs in the family: Finding dangerous drugs using drug relatedness. Presented at the Western North American Region of the International Biometric Society Annual Meeting, Los Angeles, California

2012, Shaddox TR and Suchard MA. Improving Bayesian methods in pharmacovigilance with drug hierarchies. Presented at the Joint Statistical Meetings, San Diego, California

CHAPTER 1

Introduction: the challenge of large scale observational healthcare studies

“Data Science" has found medicine. The integration of technology into medical practice and record-keeping continues to transform medical research, giving birth to interdisciplinary fields like biomedical informatics while simultaneously stimulating new, overlapping branches of biostatistics, computer science, and epidemiology. The breadth of shareholders reflects the challenges and promise of learning from observational healthcare data. Three main domains define this area. The first centers on the data: its collection and concerns surrounding its use. The second domain I will loosely call the informatics infrastructure: how the data are stored and represented. Finally, the challenge most relevant to this dissertation is modeling at scale: how we transform the data into actionable knowledge.

1.1 The data

The first question that should arise in a discussion of learning from observational healthcare data is "What added value do large repositories of medical data bring to medicine and public health?" Schneeweiss and Avorn [2005] remind us that clinical trials cannot answer all important medical questions. The marginal cost of gathering observational healthcare data is considerably smaller for the same number of patients than for traditional modalities of medical studies. Randomized trials require recruiting and maintaining patients. In contrast, observational healthcare data is collected for largely non-scientific purposes: billing or patient care. Using the data as a scientific resource is largely data repurposing.

Consequently, the size of these datasets dramatically dwarfs the datasets actively collected for scientific analysis. Alone, this difference in scale is a tremendous asset. For example, drug trials have limitations for detecting rare adverse events. Enrolling thousands of participants, such trials are underpowered for sufficiently rare events. If a drug passes clinical trials and is used by a sufficiently large population, that population can help identify rare events. A classical example of this is identifying risk for drug-induced *torsade de points* [Poluzzi et al., 2009]. This condition is both extremely rare, less than 1/100,000, and clinically significant, potentially resulting in cardiac death. Clinical trials are unable to accurately estimate the risk of a given drug for producing *torsade de points* without considerable cost, and observational data become the resource of choice [Poluzzi et al., 2009].

It is meaningful to remark on the difference between efficacy and effectiveness. Clinical trials focus on appropriately measuring efficacy, the theoretical impact of a drug on a disease [Flay et al., 2005]. However, effectiveness, how the treatment behaves in practice, may be clearer from observational studies, which looks at “real" world settings [Flay et al., 2005]. This becomes muddled when asking questions of causality, but we will try to separate the division of efficacy and effectiveness from causal inference [Imai et al., 2008]. Especially as clinical care organizations become more closely monitored for their collective decision making, having the appropriate application of knowledge to clinical action within the local setting becomes important [Barieri and Maistrello, 2009]. That is, clinical effectiveness becomes a question dependent on the local patient community.

Clinical trials may be ethically unfeasible. Clinical equipoise reflects a physician’s absence of preference among treatment choices: presented with drug A or drug B, the physician is equally fine with using either. Clinical trials rely on clinical equipoise. When it egregiously breaks down, a randomized trial would be ethically wrong. For example, it might be concerning to randomize patients to placebo when comparing the treatment effect of metformin to a new anti-hyperglycemic medication in newly diagnosed diabetic patients. There are corrections to the randomized clinical trial framework that help accommodate this, such as non-inferiority trials, but the observational data do away with this conflict, as well as far subtler, less-easily corrected ones, entirely.

Married with these benefits are several challenges [Schneeweiss and Avorn, 2005]. These challenges stem predominantly from the fact that these data are not collected for scientific study. When using observational data, the analyst will be quite far removed from the data collection process, much more so than a counterpart working on survey data, for example. This often means the analyst has less control of the data contents and may be blind to what is both measured or, critically, unmeasured in the data. For a simple example, smoking status is sparsely represented in some observational healthcare data resources. In some resources, this reflects the fact that is just not relevant for the collection agency, and for others, it reflects a system-specific incomplete recording practice [Jick et al., 2000].

Patient privacy is a serious concern that drives much legislative involvement in observational health data. Mechanically, this can be problematic for accessing datasets, but that is more annoyance than flaw. True concerns arise as de-identified datasets have greater coverage, increasing the likelihood of significant patient overlap. Linking databases becomes desirable but is structurally prevented.

Similarly, drop-out is a concern. While loss-to-follow-up is not unique to observational studies, its presence is significant. Turnover rates for insurance data can reach as high as 30% per year [Short et al., 2003, Schneeweiss and Avorn, 2005]. This can be critical when considering the data as representative of long-term observation.

Here, we will consider two main types of observational healthcare data. First, electronic health records (EHR) data consist of the data stored by care centers who deploy clinical information systems to facilitate patient care. In theory, EHR data are very complete, since they represent the data important to a clinician making decisions about patient care. A drawback of EHR data is that they are largely geographically limited. For example, an EHR dataset may reflect a single care location or a shared network of care locations. But, these networks are relatively small on a national scale, raising questions about generalizing results from single EHR systems. On the other hand, insurance claims data often have larger catch basins for the patients represented and may better represent communities historically excluded from clinical trials

[Schneeweiss and Avorn, 2005]. However, these data have less granularity, and what is represented is chosen because of financial considerations.

1.2 Informatics infrastructure

Following the highly public recall of rofecoxib and renewed concern about drug safety, the U.S. Food and Drug Administration (FDA) Amendments Act of 2007 required that the FDA use observational data in an automated, reproducible approach to identify risks from medical interventions. Prior, the standard for observational pharmacovigilance was spontaneous reports data [Bennett et al., 2007]. To help address this pressing need, the Observational Medical Outcomes Partnership (OMOP) emerged [Stang et al., 2010]. Built as a public-private partnership among the FDA, academia, data owners, and the pharmaceutical industry, OMOP sought to provide a transparent framework for comparing drug safety study methodologies and evaluating drug safety risk. The OMOP community highlights the second main component of research with observational data: the considerable informatics infrastructure that must exist to facilitate meaningful analysis.

1.2.1 The common data model

In mentioning the limitations of observational data above, Schneeweiss and Avorn [2005] list one weakness of observational data as the diverse “grammars" in which data are maintained. This concern is certainly valid, but the OMOP community, and others like them, has already taken steps toward addressing this issue. One of the achievements of the OMOP community was the creation and implementation of a transparent common data model (CDM) [Stang et al., 2010]. The CDM and the extract-transform-load (ETL) processes developed to impose the CDM ensure integrity across longitudinal observational databases (LODs) [Hartzema et al., 2013].

The goal for creating the OMOP CDM was facilitating reproducibility of study designs across datasets using different terminologies. In the absence of a CDM, a researcher must run studies tailored to database specific terminology.

Largely, this means most studies only focus on a single database. Working with multiple terminologies means developing multiple studies, and such duplication invites errors. Distinct terminologies preclude assurance that studies applied to one dataset are identical to studies applied to other datasets. With a CDM, the same analyses can evaluate different datasets without modification. It follows directly that another benefit of a CDM is including multiple data sources when performing research. This allows the consumers of research to ask relevant, meaningful questions about the generalizability of results. The OMOP community brought this issue to the forefront through their landmark paper comparing studies across multiple data sources [Madigan et al., 2013].

The CDM is structured to accommodate studies on medication [Reisinger et al., 2010]. The format of the CDM is patient-centric. For each patient, there is a unique identifier that links all relevant data tables. Distinct tables hold non-overlapping information about the patient's health. For example the Drug Era table contains spans of time during which a patient was exposed to a particular product. Many different representations for a drug exist. One could report the brand name, the generic name, or the ingredients. For the Drug Era table, the CDM records the ingredients. Thus if a patient were on empagliflozin and linagliptin, a single pill with two medications to treat diabetes mellitus, from August 1, 2005 until July 5, 2006, this would register as two separate drug eras for empagliflozin and linagliptin, each from June 19, 2005 until July 5, 2007. A drug era is a combination of individual prescriptions or drug fills. For example, if the same medication is refilled routinely at the end of its 30 day supply for 2 refills, this appears as a single 90 day drug era. OMOP uses a standard 30 day persistence window, where if a new supply of the same medication is given within 30 days of the termination of a previous supply, it is considered the same era. For example, consider a patient who takes metformin for 60 days, forgets to refill a prescription for 4 days and does not take any medication. Then on the 5th day, that patient refills the prescription and continues taking metformin for 90 days. With a 30 day persistence window, all of the medication use actions result in a single 154 day drug era. The 30 day persistence window helps buffer refill discontinuities. Other tables include a Condition Era table for diagnoses and a Person table for patient demographic data.

While this illuminates the overall format for the CDM, we still need to demonstrate how the standardization occurs. This involves two goals: 1) translating from the native data representation to the CDM representation and 2) providing accurate structure for relationships within the CDM [Reisinger et al., 2010]. The CDM achieves the first goal by maintaining a dictionary of relationships between native and CDM representation. This includes information from such disparate data terminologies as the International Classification of Diseases, Ninth Revision - Clinical Modification (ICD-9-CM), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT®), and Medical Dictionary for Regulatory Activities (MedDRA®). The CDM accomplishes the second goal by encoding the relevant "is-a" relationships that represent how concepts relate. For example, the relationship from the SNOMED-CT or Anatomical Therapeutic Chemical (ATC) hierarchy present in the CDM would allow us to see that rofecoxib, an ingredient in our Drug Era table, is a cyclooxegenase-2 (COX-2) inhibitor.

Taking a native dataset and translating it to the CDM, the ETL process, follows four steps [Reisinger et al., 2010]. First, we copy the data from the source database to the appropriate CDM structure, placing the data in appropriate tables. Next, we annotate each datum with the CDM designation. Third, we aggregate data to produce the appropriate drug eras, dependent on our selected persistence window. Fourth, we export the data, now transformed to the CDM, to a relational database.

While this concludes the data processing, it is reasonable to wonder how well such processing captures the native data representation. Overhage et al. [2012] studied this question in the context of 10 observational datasets, including both claims and EHR data. OMOP created two packages to test transformation fidelity: the Observational Source Characteristics Analysis Report (OSCAR), which produces simple summary statistics, and the Generalized Review of OSCAR Unified Checking (GROUCH), which checks for anomalous results in OSCAR [Overhage et al., 2012]. Unacceptable results that GROUCH detects include, for example, abnormally large patient ages, negative era time, and pregnancy in males, among many others. Although the ETL process required substantial teams of different skills, including informaticists and clinicians, the process itself took at most 11 days. The conversion to the CDM was largely complete, around 90% or higher, and accurate, as evaluated by

GROUCH.

1.3 Modeling at scale

Armed with a wealth of promising clinical data structured to facilitate reproducible research, the last step in learning from observational healthcare data is the transition from data to meaningful, actionable inference. This is the modeling step. Many factors enter into selecting how to model healthcare events in this space. Madigan et al. [2011] unambiguously states that the central challenge of drawing inference from observational healthcare data is confounding, where identified associations between drugs and outcomes are fallaciously given causal significance. However, while addressing confounding through fidelity in modeling the underlying biology and clinical decision making is critical, practical demands motivate some modeling choices over others. In particular, the key modifier in this field is “at scale." Our datasets consist of tens of millions of patients exposed to the full spectrum of medical products. Coping with such massive resources requires considering the computational demands.

1.3.1 Univariate studies

At the beginning of modern pharmacovigilance, the dominant approach to identifying drug-outcome pairs of interest relied on disproportionality testing [DuMouchel, 1999]. The general structure of this technique is a two-by-two table, where the table is populated with the counts of patients who took the drug or did not and had the outcome or did not. Then, some measure of “interestingness" would evaluate the table, and “interesting" drugs would call for further investigation [Madigan et al., 2011]. In part, this reflected the most widely used data, reports from physicians of events that they thought were suspicious. However, this technique also emerged in the setting of claims and EHR data.

This method satisfies our second requirement for modeling “at scale." Namely, it deals with dimensionality with ease, since the reported values are essentially just counts. Any statistical inference for these disproportionality methods builds conclusions from four numbers. The weakness of this method is dealing with confounding. For an example from [Madigan et al., 2011], consider an anti-emetic drug. Further, suppose that this drug makes patients susceptible to eye infections. If an attentive physician treats a patient on the anti-emetic with a prophylactic antibiotic, this antibiotic will appear to be strongly correlated with nausea, unless we control for the anti-emetic medication. In this case, the sacrifice of confounding adjustment for performance produces a spurious result. Frustratingly, this technique continues to be near or at the cutting edge of drug safety surveillance [Huang et al., 2014].

1.3.2 Multivariate studies

Moving beyond univariate studies to more sophisticated methods represents a “quantum jump in pharmacovigilance" [Hauben et al., 2005]. Methods that have gained considerable traction for analyzing outcomes from observational datasets include cohort, case-control, and case-crossover methods [Maclure, 1991, Rothman et al., 2008]. Of these, cohort methods remain popular for controlling for confounding variables [Schneeweiss, 2014]. Cohort methods help correct the shortcomings of the univariate disproportionality testing. However, these methods may encounter difficulties accommodating the second requirement for modeling with observational healthcare datasets: working at scale. In response, other approaches have gained popularity.

Farrington [1995] proposes the *self-controlled case series* (SCCS) method in order to estimate the relative incidence of rare drug-specific outcomes to assess vaccine safety. Risk of an event is a function of exposure-specific effects and patient-specific risks. For Farrington [1995], the exposures are vaccine delivery time intervals; for us, the exposures are medical products. For many applications, including pharmacovigilance, we are only interested in the exposure-specific effects. Estimating a per-person underlying rate of an outcome does not help identify what drugs are harmful. The SCCS model allows us to avoid estimating these unhelpful covariates, and we gain efficiency in estimating the meaningful covariates.

Simply, in these conditional methods, each patient acts as his or her own control. To this extent, we only focus on the increased rate of an outcome of interest when exposed to a drug compared to absence of exposure versus looking at the overall rate of that outcome per patient. In essence, the patient specific rate defines how many events we see for that patient, but we really just care about how that rate changes under exposure to a treatment of interest.

To explain this more rigorously, we follow the discussion of Short et al. [2011] and Farrington [1995]. We will use notation that is consistent with our later use of this model. Consider events happening as a non-homogeneous Poisson process. Let patients $i = 1, \ldots, N$ have a baseline risk e^{ϕ_i}. Define a period of observation as an era, indexed by k, and let l_{ik} be the time duration of the era. In the absence of exposure, we model the events of patient i in era k as $\text{Poisson}(l_{ik} \times e^{\phi_i})$.

Next we add in drug exposures. Let $j = 1 \ldots J$ index the potential exposures, with the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ measuring the instantaneous log relative risks of treatment exposure, and let $\boldsymbol{x}_{ik} = (x_{ik1}, \ldots, x_{ikJ})'$ where x_{ikj} indicates exposure to drug j in era k for person i. The drug exposures multiplicatively modulate the underlying instantaneous event intensity during constant drug exposure era k. The number of ADEs in era k of patient i is now $y_{ik} \sim \text{Poisson}(\lambda_{ik})$ where $\lambda_{ik} = l_{ik} \times e^{\phi_i + \mathbf{x}'_{ik}\boldsymbol{\beta}}$. Conditional on the drug exposure $\boldsymbol{x}_{ik}$, the density of y_{ik} is

$$p(y_{ik}|\boldsymbol{x}_{ik}) = \frac{e^{-\lambda_{ik}}(\lambda_{ik})^{y_{ik}}}{y_{ik}!} \tag{1.1}$$

For each person, the likelihood is

$$\begin{aligned} L_i &= p(\boldsymbol{y}_i|\boldsymbol{x}_i) \\ &= \prod_k \frac{e^{-\lambda_{ik}}(\lambda_{ik})^{y_{ik}}}{y_{ik}!} \\ &= \prod_k \frac{e^{-l_{ik} \times e^{\phi_i + \mathbf{x}'_{ik}\boldsymbol{\beta}}}(l_{ik} \times e^{\phi_i + \mathbf{x}'_{ik}\boldsymbol{\beta}})^{y_{ik}}}{y_{ik}!} \\ &= e^{-e^{\phi_i}\sum_k l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}}(e^{\phi_i})^{\sum_k y_{ik}} \prod_k \frac{(l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}})^{y_{ik}}}{y_{ik}!}. \end{aligned} \tag{1.2}$$

We condition on the total number of events $n_i = \sum_k y_{ik}$, the sufficient statistic, to avoid estimating ϕ_i. We model the events in each ear as a non-homogeneous

Poisson process, so the sum of the events per person also follows the Poisson distribution:

$$n_i|\boldsymbol{x}_i \sim \text{Poisson}(e^{\phi_i} \sum_{k}^{K_i} l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}). \tag{1.3}$$

For each person, the conditional likelihood is

$$\begin{aligned} L_i^c &= p(\boldsymbol{y}_i|\boldsymbol{x}_i, n_i) \\ &= \frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i)}{p(n_i|\boldsymbol{x}_i)} \\ &= \frac{e^{-e^{\phi_i} \sum_k l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}} (e^{\phi_i})^{\sum_k y_{ik}} \prod_k \frac{(l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}})^{y_{ik}}}{y_{ik}!}}{\frac{e^{-e^{\phi_i} \sum_k^{K_i} l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}} (e^{\phi_i} \sum_k^{K_i} l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}})^{n_i}}{n_i!}} \\ &= \frac{n_i! \prod_k (l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}})^{y_{ik}}}{y_{ik}! (\sum_k^{K_i} l_{ik} \times e^{\mathbf{x}'_{ik}\boldsymbol{\beta}})^{n_i}}. \end{aligned} \tag{1.4}$$

Thus, by conditioning on the sufficient statistics, the baseline risk falls out of the conditional likelihood of the data.

How does this conditioning statement improve modeling compared with an unconditional model? The utility from conditioning arises emerges as improved efficiency. Efficiency is a reflection of how well an estimator or design uses data to produce a reasonable result. To compare the SCCS method with the cohort method, we are interested in the relative efficiency, the ratio of the efficiency measures for each method.

Returning to Farrington [1995], we can measure the relative asymptotic efficiency of the SCCS method compared with the cohort method with the ratio of the asymptotic variance of maximum likelihood parameter estimates (MLE). To look at the asymptotic relative efficiency, consider N patients as before, each exposed to 2 eras of length l_1 and l_2. In particular, we assume no exposure happens during the first era, and we let an exposure to a single drug happen for V individuals during the second era. In the SCCS model, the log likelihood for these patients is

$$L_{sccs}(\boldsymbol{\beta}) = \sum_{n=1}^{N} \left[y_{i2}\, x_{i2}\beta - (y_{i1} + y_{i2}) \log \left(l_1 + l_2\, e^{x'_{i2}\beta} \right) \right]. \tag{1.5}$$

For $\hat{\beta}_{sccs}$, the MLE, Farrington [1995] finds

$$var(\hat{\beta}_{sccs}) = \frac{(l_1 + l_2 e^{\beta})^2}{(\sum_i y_{i1} + y_{i2})(l_1 l_2 e^{\beta})}. \tag{1.6}$$

To this, Farrington [1995] compares the unconditional cohort model. This has Poisson rate $e^{\phi}[l_1 + l_2 e^{\beta x}]$, where the patient specific rate ϕ_i is the same for all patients. Under this model, the log likelihood is

$$L_{cohort}\,(\boldsymbol{\beta}) = \sum_{n=1}^{N} \left[(y_{i1} + y_{i2})\phi + y_{i2}\beta x_i - e^{\phi}(l_1 + l_2 e^{\beta x}) \right]. \tag{1.7}$$

For $\hat{\beta}_{cohort}$, the MLE for the cohort model, Farrington [1995] finds

$$var(\hat{\beta}_{cohort}) = \frac{1}{V(l_2 e^{\phi+\beta})} \left(1 + \frac{l_2(\frac{V}{N})e^{\beta}}{l_1 + l_2(1 - \frac{V}{N})} \right). \tag{1.8}$$

We compute the efficiency as

$$\begin{aligned} E &= \frac{var(\hat{\beta}_{cohort})}{var(\hat{\beta}_{sccs})} \\ &= \frac{\frac{1}{V(l_2 e^{\phi+\beta})} \left(1 + \frac{l_2(\frac{V}{N})e^{\beta}}{l_1 + l_2(1 - \frac{V}{N})} \right)}{\frac{(l_1 + l_2 e^{\beta})^2}{(\sum_i y_{i1} + y_{i2})(l_1 l_2 e^{\beta})}} \end{aligned} \tag{1.9}$$

The expected value of $\sum_i (y_{i1} + y_{i2})$ is $Ve^{\phi}[l_1 + l_2 e^{\beta}]$ using the cohort model. Therefore, we substitute and get the asymptotic efficiency as

$$\begin{aligned} E &= \frac{\frac{1}{V(l_2 e^{\phi+\beta})} \left(1 + \frac{l_2(\frac{V}{N})e^{\beta}}{l_1 + l_2(1 - \frac{V}{N})} \right)}{\frac{(l_1 + l_2 e^{\beta})^2}{Ve^{\phi}[l_1 + l_2 e^{\beta}](l_1 l_2 e^{\beta})}} \\ &= \frac{1 + \frac{l_2 \frac{V}{N}}{l_1 + l_2(1 - \frac{V}{N})} e^{\beta}}{1 + \frac{l_2}{l_1} e^{\beta}}. \end{aligned} \tag{1.10}$$

Looking at this asymptotic efficiency, we can glean some intuition about how the efficiency varies with dataset characteristics. As the proportion of treated people $\frac{V}{N}$ declines, the relative efficiency declines. That is, for situations where the exposure use is high, we see more efficiency in the SCCS method. Furthermore, as the incidence of the outcome of interest decreases, we see increase in relative SCCS efficiency. This is particularly noteworthy, since we are frequently interested in rare outcomes. In summary, both of the situations that boost the relative efficiency of SCCS apply to the problems of interest, validating our use of conditional models in this setting.

An additional benefit of the SCCS model is that it reduces the study population. If we are interested in patients with a rare event, we just use the patients in the dataset who have ever had that event. In datasets at the scale of interest, rarely do we lack for patients. Often, the challenge is computationally managing the patients we have. By limiting ourselves to the most informative patients, we can mitigate the computational burden.

CHAPTER 2

Hierarchical modeling of multiple outcomes

Clinical trials often lack power to identify rare adverse drug events (ADEs) and therefore cannot address the threat rare ADEs pose, motivating the need for new ADE detection techniques. Emerging national patient claims and electronic health record databases have inspired post-approval early detection methods like the Bayesian self-controlled case series (BSCCS) regression model. Existing BSCCS models do not account for multiple outcomes, where pathology may be shared across different ADEs. We integrate a pathology hierarchy into the BSCCS model by developing a novel informative hierarchical prior linking outcome-specific effects. Considering shared pathology drastically increases the dimensionality of the already massive models in this field. We develop an efficient method for coping with the dimensionality expansion by reducing the hierarchical model to a form amenable to existing tools. Through a synthetic study we demonstrate decreased bias in risk estimates for drugs when using conditions with different true risk and unequal prevalence. We also examine observational data from the MarketScan Lab Results dataset, exposing the bias that results from aggregating outcomes, as previously employed to estimate risk trends of warfarin and dabigatran for intracranial hemorrhage and gastrointestinal bleeding. We further investigate the limits of our approach by using extremely rare conditions. This research demonstrates that analyzing multiple outcomes simultaneously is feasible at scale and beneficial.

2.1 Introduction

Adverse drug events (ADEs) pose a serious public health risk. While clinical trials remain the gold standard for evaluating drug safety and efficacy, the emergence of massive healthcare repositories, in the form of longitudinal observational databases (LODs), introduces a novel resource for asking and answering drug safety questions. These databases contain insurance claims and electronic health records, with time-stamped patient data that include drug exposures and diagnoses. The scale of these datasets is remarkable, with hundreds to thousands of observations on tens of millions of patients. These resources can potentially support post-approval surveillance for ADEs, where we can monitor the relative safety of drugs after they are clinically available. The development of a common data model (CDM) for LODs through the Observational Medical Outcomes Partnership (OMOP) experiment facilitates statistical methods implementation using these data to address pertinent questions about health practices, including comparative drug safety [Overhage et al., 2012]. The OMOP experiment has demonstrated the value and efficacy of competing analytical approaches [Stang et al., 2010]. While observational studies may be vulnerable to variability of study design, and the OMOP community produced the first steps toward systematic statistical evaluation of observational evidence [Madigan et al., 2014].

Commensurate with its considerable promise, analysis of LODs presents a significant statistical and computational challenge. Patients have different levels of illness and compliance that are not readily identifiable from the LODs. Observations are incomplete and inhomogeneous over time. In addition, the scale of the data creates a massive, but extremely sparse, resource. Not only are LODs massive in the number of patients recorded, they also contain the full spectrum of medical products, interventions, and diagnoses. This scale precludes many analytic approaches.

ADEs are clinical manifestations of specific pathologies. For example, hypocoagulability affects the entire body, creating a general increased risk of bleeding. However, the clinician will identify the results of hypocoagulability by the anatomic location where a bleeding event occurs. If the bleeding occurs in the brain, the diagnosis will be an intracranial hemorrhage. If the bleeding occurs in the stomach, the diagnosis will be a gastric hemorrhage. The clinician will identify the outcome but may not identify the pathology. The drug-specific effect often occurs at the level of the pathology, but the identified ADEs appear at finer granularity. Connecting outcomes and drugs without considering shared pathology ignores a crucial component of the pathophysiology.

Currently, most analytical approaches consider one outcome at a time, ignoring relationships among the outcomes. In particular, we miss an opportunity to "borrow strength" [DuMouchel, 2012] across outcomes where there is shared pathophysiology. Dealing with multiple ADE outcomes remains of critical importance to epidemiology and data mining [Thuraisingham et al., 2009, DuMouchel, 2012]. DuMouchel [2012] and Crooks et al. [2012] approach this problem by borrowing strength across outcomes to construct a multivariate logistic regression.

A common method for avoiding multiple outcomes is aggregating all the outcomes of interest into one overarching category, essentially considering different outcomes as exchangeable. Selecting which outcomes are related often follows directly from how clinicians codify diseases. For example, the International Classification of Diseases version 9 (ICD-9) code 432 represents “other and unspecified intracranial hemorrhage," of which 432.1 “subdural hemorrhage" is a subtype. Using all 432.* ICD-9 codes would capture all the subtypes of “other and unspecified intracranial hemorrhage" the ICD-9 considers, essentially aggregating all subtypes under the 432 code. The OMOP Standard Vocabulary encompasses multiple disease relationship representations, including the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) vocabulary. However, determining which outcomes are related by shared pathology need not be limited to disease codes; the discretion of a clinical expert should guide their selection.

Aggregating outcomes produces drug risk estimates that reflect a weighted average of the risk for each outcome separately. This may introduce bias into outcome-specific risks. Prevalence differences underscore this bias, with high prevalence outcomes driving risk estimates. When considering outcomes with low prevalence, we would like to combine information about them with closely related common outcomes. However, aggregating these rare outcomes with common ones overwhelms the drug-outcome specific relationship. Therefore, we would like a way to treat similar outcomes as distinct while still respecting their relatedness.

In this paper we move beyond focusing on one outcome at a time. Specifically, we seek to reduce the bias that arises when we aggregate multiple, related outcomes into one synthetic outcome. To do this, we develop a set of open-source statistical tools relying on LODs structured according to the OMOP common data model. We integrate a hierarchy of pathology and outcomes into ADE detection.

2.2 The self-controlled case series (SCCS) model

2.2.1 SCCS framework

The most common approaches to analyzing outcomes from LODs include cohort, case-control, and case-crossover methods [Maclure, 1991, Rothman et al., 2008]. However, other approaches have gained popularity in recent years. Farrington [1995] proposes the *self-controlled case series* (SCCS) method in order to estimate the relative incidence of rare drug-specific outcomes to assess vaccine safety. Simpson et al. [2013] and Suchard et al. [2013] use this model successfully in ADE detection. A significant benefit of the SCCS model is that it reduces the sample size to exposed patients experiencing at least one adverse event. Adverse event risk is a function of drug-specific effects and patient-specific risks, including underlying conditions. However, we are only interested in the drug-specific effects, and the SCCS model allows us to focus our statistical power on estimating these covariates of interest. These benefits make the SCCS model ideal for pharmacovigilance. A major limitation of the SCCS remains its formulation around one outcome at a time, a situation we will rectify by splicing our hierarchical model into an SCCS framework.

The SCCS model assumes that ADEs arise according to an inhomogeneous Poisson process. For a given LOD, let P count the number of outcome types we are considering, and let $p = 1, \dots, P$ index these outcomes. For a given drug j, let Q_j equal the number of outcomes where at least one patient who has that outcome consumed that drug. Let $j_p = 1, \dots, J_p$ index the drugs where there is at least one exposure to a patient with outcome p, such that $J_1, \dots, J_P$ count the total number of drugs observed in the exposure set for each outcome. Parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_P)'$ where $\boldsymbol{\beta}_p = (\beta_{p1}, \dots, \beta_{pJ_p})'$ measure the instantaneous, unknown, log relative risks given exposure for each drug with respect to each outcome. Under the model, let patient $i = 1, \ldots, N$ for outcome $p = 1, \ldots, P$ have a baseline risk $e^{\phi_{ip}}$. We consider drug eras as intervals of exposure over which the drugs a patient takes remains constant. Let the drug exposures multiplicatively modulate the underlying instantaneous event intensity λ_{ikp} during constant drug exposure era k.

We consider drug eras as intervals of exposure over which the drugs a patient takes remain constant. This aspect of the OMOP CDM requires special attention. We use the OMOP CDM 4 definition of a drug era. A drug era is a combination of individual drug exposures, such as individual prescription fills. For example, if the same medication is refilled routinely at the end of its 30 day supply for 3 refills, this appears as a single 90 day drug era. Our constant eras are intervals of time where patients remain on the same combination of medication. For example, consider a patient who takes drug A from July 5, 2009 through July 20, 2009 and drug B from July 10, 2009 to July 17, 2009. Three distinct drug eras emerge: one era from July 5 to July 9; another from July 10 to July 17; and the last era from July 18 to July 20.

Let K_{ip} be the total number of drug eras for person i in condition p. Following the notation of Suchard et al. [2013] and Simpson et al. [2013], the intensity arises as $\lambda_{ikp} = e^{\phi_{ip} + \boldsymbol{x}'_{ikp}\boldsymbol{\beta}_p}$, where $\boldsymbol{x}_{ikp} = (x_{ikp1}, \ldots, x_{ikpJ_p})'$ and x_{ikpj} indicates exposure to drug j in era k for outcome p. The exposure duration for exposure era k of patient i is l_{ikp}. The number of ADEs in era k of patient i for outcome p is $y_{ikp} \sim \text{Poisson}(l_{ikp} \times \lambda_{ikp})$. The SCCS method conditions on the total number of events for a particular outcome $n_{ip} = \sum_k y_{ikp}$ that a patient experiences over her total observation period. For multiple outcomes, $(n_{i1}, \ldots, n_{ip})$ remain sufficient statistics for the subject's baseline risks $(\phi_{i1}, \ldots, \phi_{ip})$. By conditioning on these statistics, the baseline risks fall out of the conditional likelihood of the data regardless of their correlation and hence greatly reduce the number of parameters to estimate:

$$\begin{aligned}\prod_{i=1}^{N}\prod_{p=1}^{P} P\left(y_{ip}|x_{ip}, n_{ip}\right) &= \prod_{i=1}^{N}\prod_{p=1}^{P} \frac{P\left(y_{ip}|x_{ip}\right)}{P\left(n_{ip}|x_{ip}\right)} \\ &\propto \prod_{i=1}^{N}\prod_{p=1}^{P}\prod_{k}^{K_{ip}} \left(\frac{e^{\boldsymbol{x}'_{ikp}\boldsymbol{\beta}_p}}{\sum_{k'}^{K_{ip}} l_{ik'p} e^{\boldsymbol{x}'_{ik'p}\boldsymbol{\beta}_p}}\right)^{y_{ikp}}.\end{aligned} \tag{2.1}$$

Taking the log of Equation (3.1) yields the log-likelihood under our model

$$L(\boldsymbol{\beta}) = \sum_{n=1}^{N} \left\{ \sum_{p=1}^{P} \left[\sum_{k=1}^{K_{ip}} \left(y_{ikp}\, \boldsymbol{x}'_{ikp} \boldsymbol{\beta}_p \right) - n_{ip} \log \left(\sum_{k=1}^{K_{ip}} l_{ikp}\, e^{\boldsymbol{x}'_{ikp} \boldsymbol{\beta}_p} \right) \right] \right\}$$

that forms only part of our objective function of interest. Specifically we work in a Bayesian framework and choose to specify a prior distribution for the covariates.

Bayesian techniques are ideal for pharmacovigilance, succinctly capturing clinical prior knowledge of drug safety, and are common in the field, as seen in Curtis et al. [2008], Madigan et al. [2011]. Furthermore, the Bayesian approach mitigates many of the challenges of massive sparse data. Simpson et al. [2013] reduce overfitting under a maximum likelihood approach by assuming a prior over the drug effect parameter vector, constructing a Bayesian SCCS. We assume *a priori* that most drugs are safe and therefore assume a prior that shrinks the parameter estimates toward 0.

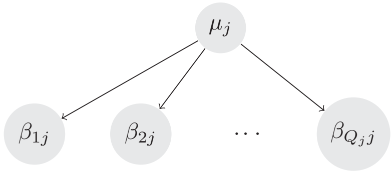
2.2.2 Disease hierarchies

To analyze a group of related outcomes, we follow DuMouchel [2012] in framing our approach as a hierarchical multivariate regression, where the specific outcomes are related under their shared pathology. Each adverse event has a separate representation of each shared drug, a drug-outcome effect estimate. We rely on our Bayesian perspective and project that idea onto multiple ADE outcomes by extending our prior.

In the original Bayesian SCCS formulation applied to LODs, there can exist upwards of $J_p \sim 10,000$ drug covariates. Multiple outcomes exacerbate this extreme dimensionality. Namely, we need to compute $\mathcal{J} = \sum_{p=1}^{P} J_p$ covariates, roughly P-fold more covariates. To cope with this ultra high dimensionality, we model the effects of the same drug across outcomes hierarchically. We represent each drug-outcome effect as inheriting from a drug-pathology effect. We extend the prior structure of the original Bayesian SCCS model by using a hierarchical prior that shares information across regression coefficients $(\beta_{1j}, \ldots, \beta_{Q_j j})$ that measure the association of a single drug j across all Q_j outcomes where drug j

appears in the records. The drug-level precision is τ_d, and the pathology-level precision is τ_p.

Not all drugs need be present across all outcomes. Therefore, we scale the prior precision for each drug by the number of outcomes in which the drug appears as a non-zero covariate. For example, if drug A appears among the patients with intracranial hemorrhage and gastric hemorrhage, while drug B appears only among patients with gastric hemorrhage, we seek to compensate for this mismatch by scaling the universal drug-level precision when approaching each outcome specific risk estimate. Specifically, we model



$$\begin{aligned} \mu_j &\sim \text{Normal}\left(0, \tau_p\right), \text{ and} \\ \beta_{1j}, \ldots, \beta_{Q_j j} &\sim \text{Normal}\left(\mu_j, Q_j \cdot \tau_d\right). \end{aligned} \tag{2.2}$$

2.2.3 Computational swindle

As described, the hierarchical model imposes greater dimensionality, a more cumbersome log-likelihood, and a host of new parameters to track, suggesting that we will require new inference equipment that scales for LODs. However, a redefinition of parameters demonstrates that our more complex model easily compresses into a form that allows for inference with the existing high performance SCCS tools of Suchard et al. [2013]. We concatenate outcome specific event counts vectors $\tilde{\boldsymbol{y}} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_P)'$ and time of exposure vectors $\tilde{\boldsymbol{l}} = (\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_P)$ into new vectors representing the adverse events and exposure times across all outcomes.

In practice, we take our data, a set of event counts and drug exposures, for each outcome and add an outcome-specific tag to each of the drug exposures. That is, each drug exposure now has an associated outcome. For example, if we look at bleeding events, with outcomes intracranial hemorrhage and gastric hemorrhage, and drug warfarin, a covariate would be warfarin-intracranial hemorrhage or warfarin-gastric hemorrhage. Considering $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_P)$, covariates for the same event are consecutive. We construct a new design matrix $\tilde{\boldsymbol{X}}$,

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{X}_P \end{bmatrix}.$$

This design matrix is necessarily block diagonal, since the outcome-specific covariates are not represented in other outcomes. For example, the warfarin-intracranial hemorrhage covariate is not present among the data on patients who have gastric hemorrhage events. Given this structure, the resulting log-likelihood is

$$L\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{N} \left[\sum_{k=1}^{\mathcal{K}_i} \left(\tilde{y}_{ik}\, \tilde{x}'_{ik} \tilde{\boldsymbol{\beta}} \right) - n_i \log \left(\sum_{k=1}^{\mathcal{K}_i} \tilde{l}_{ik}\, e^{\tilde{x}'_{ik} \tilde{\boldsymbol{\beta}}} \right) \right]. \tag{2.3}$$

Under this reindexed representation, log-likelihood (2.3) matches the expression in Suchard et al. [2013], enabling us to recycle existing computational infrastructure. Furthermore, each $\boldsymbol{X}_p$ is extremely sparse, and the computational approach of Suchard et al. [2013] efficiently represents and computes over sparse systems. While creating $\tilde{\boldsymbol{X}}$ increases the dimensionality, it is a sparse expansion, mitigating the computational demand. Thus, we can leverage the extant sparse computing solutions to evaluate this more sophisticated model, without drastically increasing the computational demand.

2.2.4 Maximum *a posteriori* estimation using cyclic coordinate descent

Given the reformulation, the changes to the univariate Bayesian SCCS framework remain in the prior. For notation, we consider the set G_j of covariates representing the same drug across all conditions we consider. The cardinality of G_j is Q_j. Let $G_{j\{p\}}$ be this set excluding β_{pj}. We consider the induced prior distribution,

$$p(\beta_{G_j}) = \int p(\beta_{G_j} | \mu_j) p(\mu_j) \mathrm{d}\mu_j. \tag{2.4}$$

Taking the log of the integrand and recalling that all coefficients are independent given the pathology effect yields:

$$
\begin{aligned}
\log\left[p(\mu_j)\right] + \log\left[p(\beta_{G_j}|\mu_j)\right] = &\left[f_1(\tau_p) - \frac{\tau_p}{2}(\mu_j - 0)^2\right] + \\
&\left[f_2(\tau_d) - \frac{Q_j\tau_d}{2}\sum_{g\in G_j}(\beta_g - \mu_j)^2\right] \\
=&f_3(\tau_p, \tau_d) - \frac{\tau_p}{2}(\mu_j - 0)^2 - \frac{Q_j\tau_d}{2}\sum_{g\in G_j}(\beta_g - \mu_j)^2 \\
=&f_3(\tau_p, \tau_d) - \frac{1}{2}\left[(Q_j^2\tau_d + \tau_p)\mu_j^2 - 2Q_j\tau_d\mu_j\sum_{g\in G_j}\beta_g\right. \\
&\left. + Q_j\tau_d\sum_{g\in G_j}\beta_g^2\right].
\end{aligned}
$$

In this construction, $f_1(\tau_p)$, $f_2(\tau_d)$, and $f_3(\tau_p, \tau_d)$ are constants with respect to μ_j and G_j employed to simplify notation.

Completing the square to perform the integral returns

$$
\log\left[p(\beta_{G_j})\right] = f_4(\tau_p, \tau_d) - \frac{1}{2}\left[Q_j\tau_d(\sum_{g\in G_j}\beta_g^2) - \frac{(Q_j\tau_d\sum_{g\in G_j}\beta_g)^2}{Q_j^2\tau_d + \tau_p}\right], \qquad (2.5)
$$

where $f_4(\tau_p, \tau_d)$ is a constant with respect to μ_j and G_j that remains after integrating over μ_j.

The implementation of Suchard et al. [2013] uses cyclic coordinate descent (CCD) to find the maximum *a posteriori* (MAP) estimates through optimizing the model log posterior $P(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \log[p(\boldsymbol{\beta})]$. Our approach amounts to regularized regression, for which CCD has been heavily employed [Friedman et al., 2010, Wu and Lange, 2008]. CCD circumvents the need to invert the full Hessian at each step [Wu et al., 2009b]. At each CCD iteration, the updates are a function of the log-likelihood gradient $\partial L/\partial\beta_{pj}$ and Hessian $\partial^2 L/\partial\beta_{pj}^2$

as well as the penalty gradient $\partial \log[p(\boldsymbol{\beta})]/\partial\beta_{pj}$ and Hessian $\partial^2 \log[p(\boldsymbol{\beta})]/\partial\beta_{pj}^2$. A single Newton step is taken along each coordinate and proves extremely efficient when $\boldsymbol{X}$ is sparse [Genkin et al., 2007, Suchard et al., 2013].

Working in the CCD framework, we require the gradient and the Hessian contributions to the log-likelihood and log-priors. Fortunately, the log-likelihood remains unchanged using our computational swindle. However, the penalty component does change under the hierarchical model, with both the gradient and the Hessian a function of the pathology precision. The forms of the penalty components in the Newton steps are

$$\begin{aligned} \frac{\partial \log(p(\beta_{pj}|\beta_{G_{j\{p\}}}))}{\partial\beta_{pj}} &= -Q_j\tau_d\beta_{pj} + \frac{(Q_j\tau_d)^2(\sum_{g\in G_j}\beta_g)}{Q_j^2\tau_d+\tau_p} \text{ and} \\ \frac{\partial^2 \log(p(\beta_{pj}|\beta_{G_{j\{p\}}})}{\partial{\beta_{pj}}^2} &= -Q_j\tau_d + \frac{(Q_j\tau_d)^2}{Q_j^2\tau_d+\tau_p}. \end{aligned} \tag{2.6}$$

2.2.5 Hyperparameter selection:

We use cross-validation based on the predictive log-likelihood of the hold-out set to select the hyperparameters τ_p and τ_d. Suchard et al. [2013] use a log-scale grid search that is computationally expensive even with only a single parameter. When we add a second parameter, this method becomes impractically slow. The additional parameter τ_p increases overall computing cost by an order of magnitude. However, it remains desirable to use cross-validation to select both τ_d and τ_p.

To help overcome this burden, we turn to Genkin et al. [2007] in implementing an "autosearch" for hyperparameter selection. We start with an initial guess and then increase or decrease our guess by one log unit until we have bracketed the maximum of the hold-out set predicted log-likelihood. Then we compute a quadratic approximation to the predicted log-likelihood. The maximum of this approximate surface becomes our estimate. To find both hyperparameters, we alternate between them, fixing one and finding the conditional maximum of the other, and then fixing to that new conditional solution and finding the conditional maximum of the other. We continue this process until both previous and proposed hyperparameters are within an order of magnitude.

We prefer using this flexible tolerance method to a fixed tolerance method, in which finding the appropriate fixed tolerance would be difficult considering the log-scale of the search space.

2.3 Demonstration:

2.3.1 Synthetic study: biased risk estimates

To evaluate the bias that arises when using aggregated outcomes, we simulate a small dataset with three conditions of interest. For the first and third conditions, the prevalence of these diagnoses is extremely low, with only 20 and 10 patients having these conditions, respectively. For the second condition, the prevalence is much higher, with 1000 patients present in our hypothetical dataset. We expose these synthetic patient groups to 10 drugs. Two drugs are positively associated with all conditions. However, the risk for these two drugs varies drastically among the three groups. In particular, the two dangerous drugs present a log relative risk of 0.5 for the first, rare condition, a log relative risk of 1 for the second common condition, and a log relative risk of 2 for the third, rare condition.

In our simulations, we first draw a patient-specific underlying risk from a Normal(-1,0.5) distribution. Then, for each patient, we uniformly select between 1 and 10 observations, or drug exposure eras, as well as an observation length per observation. In each observation, we assign between 1 and 10 drugs to the patient. For each drug, we know the log relative risk for the given event. Armed with the underlying risk rate as well as the drugs per observation, we compute the overall risk rate for each observation and draw from a Poisson distribution with that intensity to get the event count during that observation.

We compare the marginal estimates of the relative risks in both the aggregated data situation and using our hierarchical model. We first run our analysis considering all conditions exchangeable, extracting one risk estimate per drug. Effectively, when we aggregate data, the log relative risk among these three populations becomes a weighted average risk estimate. In Figure (2.1 a) we see that the analysis of the aggregated data slightly underestimates the log

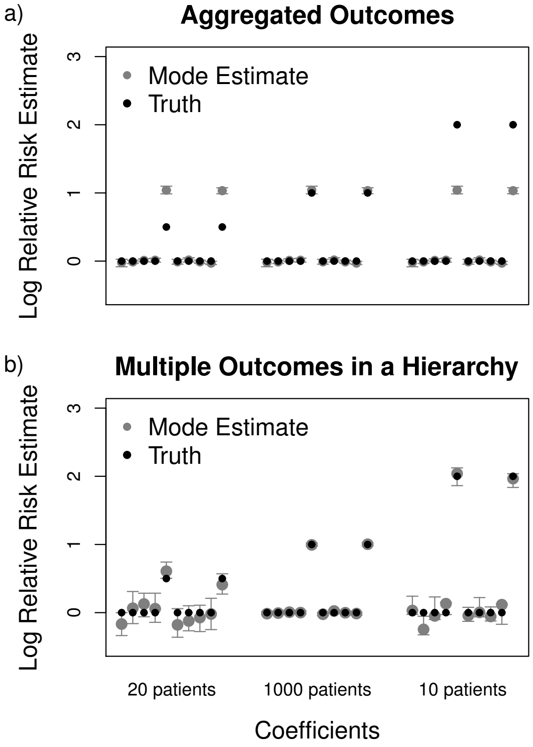


Figure 2.1: *Mode estimates and 95% bootstrap confidence intervals (gray) of the log relative risk for each drug and their simulated relative risk (black) across two conditions with different prevalence. The first 10 covariates represent the estimates from one condition with a prevalence of 20 patients; the second 10 represent estimates from the condition with high prevalence, affecting 1000 patients; and the last 10 covariates represent a second condition with low prevalence, affecting 10 patients. Using the multiple outcomes in an aggregated approach (a) produces less appropriate estimates than the hierarchical outcomes approach (b).*

relative risk of the dangerous drugs in the large population. In the 20 patient and 10 patient populations, the method seriously overestimates and underestimates, respectively, the log relative risk of the dangerous drugs. In contrast, the estimates from modeling these outcomes together under a hierarchical structure avoid this problem. Separate risk estimates for each drug-outcome pairs demonstrate much less bias, as seen in Figure (2.1 b).

We compare the model fitting times for each of these datasets, including cross-validation and bootstrapping with 200 replicates. For the cross-validation, we averaged the predicted log-likelihood over 6 permutations of the 10-fold sampling of the data. Fitting the aggregated dataset took 5 seconds, and the cross-validation variance was 0.1. Using the autosearch cross-validation method, fitting the hierarchical model took 9 seconds. We also fit the model using the grid search cross-validation method. Specifically, we used a 10 by 10 grid ranging from 10^{-4} to 10^{5} for both τ_p^{-1} and τ_d^{-1} . Using this grid, fitting the model took 32 seconds. The results from the autosearch, with starting estimates of 100, produced estimates of τ_p^{-1} and τ_d^{-1} at 1.1 and 2.2; the results from the grid search produced estimates of τ_p^{-1} and τ_d^{-1} at 100 and 1.

The difference between the estimates for τ_p^{-1} from the autosearch method and the grid-search method is noteworthy. The autosearch method finds a value beyond a grid point of 100. This results from two effects. First, both the autosearch and grid-search estimates may be sensitive to fitting parameter choices, like the number of permutations over which to average. This reflects the relatively flat topology of the predictive log-likelihood in this small dataset, where chance selections of data for cross-validation can move our perceived apex. We remedy this partly by averaging over multiple data permutations. Second, this difference underscores the inability of the grid method to adjust resolution as needed. The grid-search method is bound by our decision of grid size. Resolving the method using a finer grid is computationally daunting. The autosearch method avoids this problem, adjusting resolution as needed without the computational tax. However, the difference we see between the search methods in this case fails to appreciably change the estimated relative risks, with no risk estimates changing by more than 0.015. This result highlights the stability of our risk estimates to different hyperprior estimates.

2.3.2 Real world study: warfarin and dabigatran

The standard for outpatient anticoagulation is warfarin, an inhibitor of vitamin K metabolism. Clinically, warfarin is difficult to use, requiring frequent laboratory tests to identify its sensitive, patient-specific dosing. Alternatives to warfarin present an opportunity for improving anticoagulation care. In 2009, a randomized, controlled, noninferiority trial suggested that dabigatran etexilate has a comparable treatment effect to warfarin [Connolly et al., 2009]. Furthermore, the manufacturer claims that dabigatran requires less clinical attention than warfarin to find the appropriate dose. Although this trial also found grossly similar risk profiles for dabigatran and warfarin, there were notable differences. In particular, warfarin posed a greater overall risk of major bleeding. However, dabigatran posed a significantly elevated risk of gastrointestinal hemorrhage (GIH). Among the worst outcomes for patients on anticoagulation therapy with warfarin is intracranial hemorrhage (ICH). The rate of this ADE among dabigatran patients was one third that of warfarin patients. Thus, for one ADE, dabigatran appears to increase risk; for another, it appears to be safer. Many concerns about this trial have surfaced [Charlton and Redberg, 2014]. New events of interest from the trial emerged later [Connolly et al., 2010]. Reilly et al. [2014] produced better dose-risk trade-off results. Subsequent clinical trials have reexamined the risk of major bleeding events. The results of these trials are equally inconclusive, with greater transfusion needs among dabigatran treated patients counterbalanced by lower intensive care stay and lower mortality [Majeed et al., 2013].

We contribute to this debate by considering a real world equivalent of the simulated study above. We want to use our hierarchical model to tease out the risk profiles for both warfarin and dabigatran while reflecting the shared pathology of bleeding events. Thus, we consider each of these outcomes under our hierarchical model. Furthermore, we explore what would happen if we aggregated these data, considering GIH and ICH exchangeable.

To perform these studies, we examine the MarketScan Lab Results (MSLR) dataset, maintained by the Reagan-Udall Foundation Innovation in Medical Evidence Development and Surveillance project. This dataset comprises 1.5 million patient lives. We depend on the OMOP common data model version 4

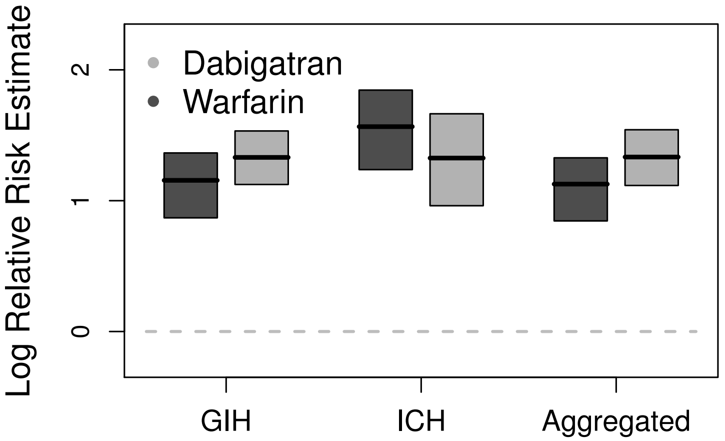


Figure 2.2: *Mode estimates and 95% bootstrap confidence intervals for the effect of dabigatran (light gray) and warfarin (dark gray) on gastrointestinal hemorrhage (GIH) and intracranial hemorrhage (ICH), compared to an aggregated outcome where GIH and ICH are exchangeable.*

for representation of concepts of interest. To examine GIH and ICH, we select all patients who experienced a diagnosis that the OMOP common data model version 4 considered a subset of GIH or ICH. There are 37,909 patients who had GIH and 2,893 patients who had ICH.

Figure (2.2) demonstrates our results. Grossly, three trends appear. First, we see that warfarin presents a lower risk for GIH than dabigatran. Second, this risk pattern reverses for ICH. This replicates trends previously found in the literature. Third, we see that considering these outcomes as exchangeable seriously masks the ICH estimates. The larger population of GIH patients overwhelms the analysis.

We again consider the computation time for each analysis, including cross-validation and bootstrapping. We used 200 replicates for the bootstrapping and averaged over 20 permutations of the cross-validation sampling data. Analyzing the GIH and ICH datasets independently took 124 and 9 seconds producing single variance estimates of 5.28 and 29.06 using one dimensional autosearch with a starting value of 0.1. Analyzing the aggregated dataset using one dimensional autosearch required 111 seconds and produced a single variance estimate of 1.1. Under the hierarchical model, the 10 by 10 log grid-search approach with a range of 10^{-3} to 10^{6} took 4735 seconds and produced estimates of τ_p^{-1} and τ_d^{-1} of 1 and 0.1. Using the two dimensional autosearch approach with an initial value of 0.001 took 2163 seconds and produced estimates of τ_p^{-1} and τ_d^{-1} of 4.15 and 0.18.

2.3.3 Real world study: extreme prevalence differences

In some cases, we want to evaluate the risk of extremely rare events, which may contain very little information about each drug risk pair. To explore what happens in this situation, we return to the MarketScan Lab Results (MSLR) dataset. Specifically we focus on two conditions: chronic gastrojejunal ulcer with hemorrhage and obstruction (CGJUHO) and vomiting blood (VB). Both of these diagnoses inherit from the OMOP common data model version 4 representation of upper gastrointestinal bleeding. We produce risk estimates from modeling these two categories as exchangeable, and we contrast our results when treating these two categories hierarchically. To construct our patient population, we select all patients who have had either of those diagnoses delivered in an inpatient, emergency department, or outpatient setting. There are only 24 patients with CGJUHO; there are 16,062 patients with VB. We consider the entire spectrum of drugs for both conditions.

Using 10-fold cross-validation with the predictive log likelihood averaged over two permutations and the one dimensional autosearch with an initial value of 0.1, analyzing the CGJUHO data alone produces a single prior variance estimate of 0.060 in 4 seconds, and analyzing the VB data produces a single prior variance estimate of 0.0076 in 200 seconds. The aggregated model required 200 seconds to find the point estimates, with a variance estimate of 0.0077. Under the hierarchical model, using a 10 by 10 log-scale grid of variance values ranging from 10^{-8} to 10^{1}, we find τ_p^{-1} and τ_d^{-1} maximize the predicted log-likelihood at 0.01 and 0.0001, respectively. Using the autosearch method, we find the optimal τ_p^{-1} and τ_d^{-1} to be 0.019 and 0.00017, respectively. The autosearch required 10,500 seconds; the grid search required 209,500 seconds.

Although we consider all drugs for each condition of interest, it is most interesting to look at the drugs that are present among both the set of patients with CGJUHO and the set of patients with VB. The 288 drugs that fit this criterion have non-trivial hierarchies. From Figure (2.3), we see that under the hierarchical model, the condition-specific risk estimates are very close. Furthermore, the estimates under the hierarchical model are very close to those under the aggregated model.

Ostensibly, this result undercuts the purpose of the hierarchical modeling. However, there are notable differences between this study and both the previous simulated study and the warfarin and dabigatran study. In this case, CGJUHO had drastically fewer patients than VB. Given the stark contrast in prevalences, it is reasonable for the very common condition to dominate the risk estimates of both the hierarchical and aggregated models. This suggests that the hierarchical model will correct for risk estimate bias as long as the prevalence differences between two conditions are not extreme. But, in the case of extreme prevalence differences, the results will be similar to aggregating the data. While the greedy iterative two dimensional autosearch approach greatly reduces computational time relative to the exhaustive search, it is still faster to compute a single hyperprior. Therefore, the differences in prevalence should guide the user in determining whether using the hierarchical model is warranted in her analysis.

2.4 Discussion

In this work, we have developed a novel hierarchical framework for analyzing multiple outcomes in the setting of massive observational data. We have demonstrated that we can easily restructure this framework to leverage extant inference tools that mitigate the dimensional explosion of analyzing multiple outcomes. Furthermore, we have shown the value of such a framework in better discrimination of dangerous drugs and in better risk identification in small populations.

There are challenges in working with observational data [Ryan, 2013]. Inter-database variation in reported risk estimates can be considerable [Madigan et al., 2013]. Bias in the recording of the data percolates through all analyses. Assumptions regarding the uniformity of treatment and diagnosis decisions among physicians are almost certainly incorrect. The time-invariant risk assumption underlying the SCCS model is almost certainly false for some drug and disease pairs.

However, the quantity of data from observational healthcare datasets will not decrease, and the promise of these data remains strong. One hope for success in this field is to channel the information present in these databases into a framework that optimally allows for signal detection and noise reduction. One method for achieving this goal effectively is to integrate more biological and medical knowledge into the models. The simple hierarchical model of disease, which matches both disease biology and clinical perspectives of disease, is one modest example of such structural knowledge motivating advances in modeling.

In the future, hierarchical modeling can extend beyond diseases. Drugs also follow a natural hierarchical structure. Physicians and pharmacologists use drug classification heavily to group medications with similar modes of action together. These classification systems form a natural framework for understanding drug risk. The post-approval withdrawal of Vioxx (rofecoxib) has been one of the highest profile cases of a drug with insidious side effects. The medical community did not fully appreciate the cardiac effects of rofecoxib until after the drug had been released to the market. It is thought that the entire class of COX-2 inhibitors puts patients at risk for cardiovascular events [Cannon and Cannon, 2012]. While traditional NSAIDs inhibit COX-1 and COX-2, COX-2 selective inhibitors have negligible effects on COX-1. One could consider the hierarchical structure of the drugs following a similar model as suggested here. Each of the drugs could inherit a class-specific risk. For example, all of the COX-2 inhibitors would share a greater risk for MI than the COX-1 inhibitors. This would allow the model to capture class specific effects that are currently inefficiently estimated independently.

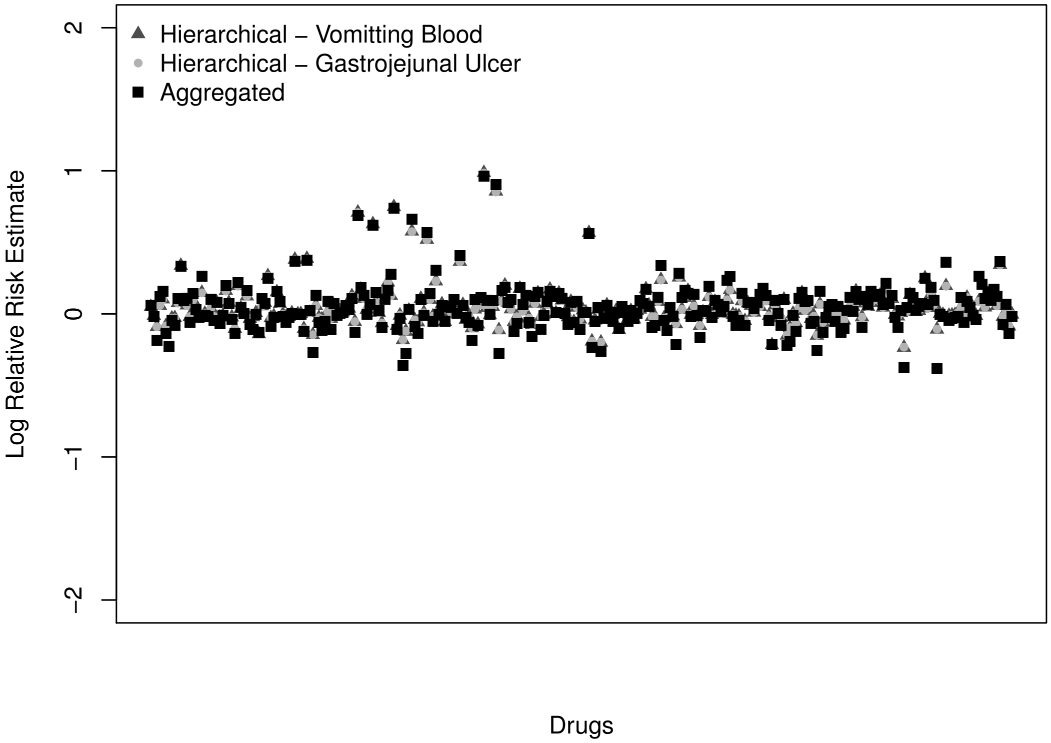


Figure 2.3: *Mode estimates of the log relative risk for each drug for a common, rare, or aggregated outcome. The common outcome is vomiting blood (VB), dark gray triangles. The rare outcome is chronic gastrojejunal ulcer with hemorrhage and obstruction (CGJUHO), light gray circles. The aggregated outcome is CGJUHO or VB, black squares. The estimates for CGJUHO and VB rely on the hierarchical structure.*

CHAPTER 3

Minorization-maximization (MM) methods for the SCCS model

Adverse drug events (ADEs) remain a serious public health risk. Identifying dangerous drugs from emerging national patient claims and electronic health record databases is a non-trivial statistical challenge. Specifically, fitting models in the context of datasets with tens of thousands of covariates and millions of observations always perches on the edge between computationally expensive and intractable. New techniques for optimization in this setting add to the arsenal of strategies that can push sophisticated, meaningful modeling toward feasibility. Here, we develop a Minorization-Maximization (MM) algorithm in the context of the Bayesian self-controlled case series (SCCS) regression model. We take two minorization transformations of the SCCS likelihood to produce parameter separation in the surrogate. Although this increases in the number of iterations required for convergence, it transforms a sequential iterative algorithm into a parallel iterative algorithm. Looking to an observational dataset examining the relationship of diclofenac and gastrointestinal bleeding with 940 drug exposure covariates and more than 5.5 million distinct drug eras, we find that parallel processing with the decoupled Newton steps improves model fitting 10 fold. We further demonstrate how acceleration with augmented Newton steps and quasi-Newton approximation can improve speedup to 17 and 28 fold above the sequential algorithm. These results underscore the value MM algorithms can bring to high-dimensional regression problems in setting of massive observational data.

3.1 Introduction

Questions of drug safety and comparative effectiveness hold considerable interest for the medical community and present novel challenges for the statistical community. While randomized controlled trials remain the gold standard, the emergence of massive healthcare data repositories presents a new setting in which to learn about drug exposures [Stang et al., 2010]. These data resources are often in the form of longitudinal observational databases (LODs), with millions of patients represented in insurance claims and electronic health records. These resources have the scope and diversity to identify rare events and to address medical product use 'in the wild.' Here we will frame our work in terms of adverse events (AEs).

Traditionally, learning about drug exposure risks relies on cohort, case-control, and case-crossover methods [Maclure, 1991, Rothman et al., 2008]. The *self-controlled case series* (SCCS) designed by Farrington [1995] has gained popularity in recent years [Simpson et al., 2013] and [Suchard et al., 2013]. Conditioning on the presence of at least one adverse event for each subject reduces the sample size to exposed patients and eliminates questions of appropriate case and control matching criteria. Additionally, under this conditioning argument, the patient-specific risks, including underlying conditions, disappear, reducing the problem to estimating drug-specific effects.

Analysis of LODs presents a significant computational challenge. These data track millions of patient observations including thousands of medical products, a massive and extremely sparse resource. Learning about associations between all products and specific AEs, while controlling for simultaneous exposures, is a compelling goal. Generalized linear models (GLMs) with unknown parameter regularization or Bayesian priors provide a fruitful and popular framework [Madigan et al. 2011]. However, naive implementation to find maximum *a posteriori* (MAP) point-estimates, often with multivariate Newton's method slows to a crawl with millions of outcomes and thousands of predictors. Within the SCCS model, mode finding remains the computational bottleneck. Simpson et al. [2013] and Suchard et al. [2013] avoid the taxing high-dimensional matrix inversion component of Newton's method and implement mode finding through cyclic coordinate descent (CCD). This is a standard strategy to mode finding in regularized regression [Friedman et al., 2010, Wu and Lange, 2008]. One of the drawbacks of this optimization strategy is that the Newton steps are inherently serial.

In this paper, we develop a method to decouple the CCD Newton steps, and, by doing so, create a parallelizable optimization procedure. We do this by leveraging the minorization-maximixation (MM) algorithm [Hunter and Lange, 2000]. The MM algorithm replaces directly evaluating the objective function of interest with computing over a surrogate function. When using the MM algorithm, the objective function of interest often satisfies some property that the original objective function did not. For us, this property is local independence of the covariates. We also demostrate two methods for accelerating optimization with this MM algorithm. We compare these strategies by looking at an observational healthcare dataset with 5.5 million drug eras and 940 distinct exposures, focusing on the risk of the painkiller diclofenac on gastrointestinal bleeding.

3.2 Methods

3.2.1 SCCS

SCCS is a cases-only design, where each individual controls for her own exposure, removing individual-specific effects. The method compares AE rates between exposed and unexposed time-intervals . The SCCS model assumes that AEs arise according to an inhomogeneous Poisson process. For $j = 1 \dots J$ drugs under consideration, the parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)'$ measure the instantaneous, unknown, log relative risks given exposure. Under the model, let patient $i = 1, \dots, N$ have a baseline risk e^{ϕ_i}. We consider drug eras as intervals of exposure over which the drugs a patient takes remains constant. Let the drug exposures multiplicatively modulate the underlying instantaneous event intensity λ_{ik} during constant drug exposure era k. That is, following the notation of Suchard et al. [2013] and Simpson et al. [2013], the intensity arises as $\lambda_{ik} = e^{\phi_i + \boldsymbol{x}'_{ik}\boldsymbol{\beta}}$, where $\boldsymbol{x}_{ik} = (x_{ik1}, \dots, x_{ikJ})'$ and x_{ikj} indicates exposure to drug j in era k . The exposure duration for exposure era k of patient i is l_{ik}. The number of AEs in era k of patient i is $y_{ik} \sim \text{Poisson}(l_{ik} \times \lambda_{ik})$. The

SCCS method conditions on the total number of events for a particular outcome $n_i = \sum_k y_{ik}$ that a patient experiences over her total observation period. By conditioning on these statistics, the baseline risk falls out of the conditional likelihood of the data and greatly reduces the number of parameters to estimate:

$$\prod_{i=1}^{N} P\left(y_i|x_i, n_i\right) = \prod_{i=1}^{N} \frac{P\left(y_i|x_i\right)}{P\left(n_i|x_i\right)} \propto \prod_{i=1}^{N} \prod_{k}^{K_i} \left(\frac{e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}}{\sum_{k'}^{K_i} l_{ik'} e^{\boldsymbol{x}'_{ik'}\boldsymbol{\beta}}} \right)^{y_{ik}}. \tag{3.1}$$

Taking the log of Equation (3.1) yields the log-likelihood under our model

$$L\left(\boldsymbol{\beta}\right) = \sum_{n=1}^{N} \left\{ \left[\sum_{k=1}^{K_i} \left(y_{ik}\, \boldsymbol{x}'_{ik}\boldsymbol{\beta} \right) - n_i \log \left(\sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}} \right) \right] \right\}. \tag{3.2}$$

Bayesian techniques are ideal for pharmacovigilance, succinctly capturing clinical prior knowledge of drug safety, and are common in the field, as seen in Curtis et al. [2008], Madigan et al. [2011]. Furthermore, the Bayesian approach mitigates many of the challenges of massive sparse data. Simpson et al. [2013] reduce overfitting under a maximum likelihood approach by assuming a prior over the drug effect parameter vector, constructing a Bayesian SCCS. We assume *a priori* that most drugs are safe and therefore assume a prior that shrinks the parameter estimates toward 0. Thus, Equation (3.2) forms only part of our objective function of interest. For each covariate, we have

$$\beta_j \sim \text{Normal}\left(0, \tau\right) \tag{3.3}$$

for precision τ. Equivalently, this transforms our optimization problem into a penalized regression problem, where we employ an L_2 norm. For a deeper understanding of the connections between penalized regression and our Bayesian formulation, Kyung et al. [2010] is a notable resource.

The implementation of Suchard et al. [2013] uses cyclic coordinate descent (CCD) to find the maximum *a posteriori* (MAP) estimates through optimizing the model log posterior $P(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \log[p(\boldsymbol{\beta})]$. The common idea behind CCD algorithms is to update $\boldsymbol{\beta}$ by cycling through β_j. Each update is a function of the unidirectional log posterior gradient $\frac{\partial P(\boldsymbol{\beta})}{\partial \beta_j}$ and Hessian $\frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_j^2}$. This approach is widely used for regularized regression [Friedman et al., 2010, Wu and Lange, 2008]. Preference for the CCD algorithm over traditional multivariate Newton's method stems from avoiding the Hessian inversion [Wu et al., 2009a].

Within each univariate Newton step, there are choices for the size of one-directional step. Rather than iterate one-dimensional updates to convergence within a cycle, many prefer taking a single Newton step [Genkin et al., 2007, Wu and Lange, 2008, Zhang and Oles, 2001]. Here, we use the implementation of Simpson et al. [2013] that follows from Genkin et al. [2007], Zhang and Oles [2001]:

$$\Delta\beta_j = -\frac{\frac{\partial}{\partial\beta_l}(L(\boldsymbol{\beta}) + \log[p(\boldsymbol{\beta})])}{\frac{\partial^2}{\partial{\beta_l}^2}(L(\boldsymbol{\beta}) + \log[p(\boldsymbol{\beta})])}. \tag{3.4}$$

We reiterate the fitting procedure from Suchard et al. [2013] in Algorithm 1, with a terser representation to highlight the structure we will contrast with subsequent algorithms. Following Genkin et al. [2007] and Suchard et al. [2013] we declare convergence when the sum of the absolute change in $\boldsymbol{X\beta}$ from successive iterations falls below 1×10^{-6}.

Algorithm 1 Cyclic coordinate descent (CCD) algorithm for fitting the Bayesian self-controlled case series model. This highlights the serial nature of the algorithm. In particular, we see $\boldsymbol{\beta}$, the J-dimensional vector of regression coefficients, as both the target over which we wish to maximize the log-posterior and the focus of our serial updates.

```
Initialize: $\boldsymbol{\beta} = 0$
while $\boldsymbol{X\beta}$ has not converged do
  for $j \in \{1, \ldots, J\}$ do
    Compute $\frac{\partial}{\partial\beta_j}(L(\boldsymbol{\beta}) + \log[p(\boldsymbol{\beta})])$ and $\frac{\partial^2}{\partial{\beta_j}^2}(L(\boldsymbol{\beta}) + \log[p(\boldsymbol{\beta})])$
    Update $\beta_j = \beta_j + \Delta\beta_j$
    Update $\boldsymbol{X\beta}$ and $\sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}\ \forall i$
  end for
end while
```

3.2.2 MM algorithm: a philosophy

The popular expectation-maximization (EM) algorithm is a critical tool for situations where closed-form score equations in maximum likelihood estimation are absent [Dempster et al., 1977]. Statisticians have come to realize that the EM algorithm is a special case of the broader class MM (minorization-maximization or majorization-minimization) algorithms [Lange et al., 2000, Hunter and Lange, 2004, Wu and Lange, 2008].

The MM algorithm is not a recipe for solving an optimization problem, but a framework for constructing algorithms. The main idea of the MM approach is to avoid maximizing a difficult function by working with an easier surrogate function. Consider an objective function $f(\boldsymbol{x})$ which we want to maximize. We seek a surrogate function that will minorize f. The surrogate function of interest $g(\boldsymbol{x}|\boldsymbol{x}_m)$ minorizes $f(\boldsymbol{x})$ if it shares the value of $f(\boldsymbol{x}_m)$ and for all other $\boldsymbol{x}$, g is below f. That is, we require

$$\begin{aligned} f(\boldsymbol{x}_m) &= g(\boldsymbol{x}_m|\boldsymbol{x}_m), \text{ and} \\ f(\boldsymbol{x}) &\geq g(\boldsymbol{x}|\boldsymbol{x}_m), \boldsymbol{x} \neq \boldsymbol{x}_m. \end{aligned} \tag{3.5}$$

In the MM algorithm, the goal is to move the surrogate function uphill, relocate the point of tangency, and recompute the surrogate function. For many MM approaches, this means actually maximizing the surrogate function

$$\boldsymbol{x}_{m+1} = argmax(g(\boldsymbol{x}|\boldsymbol{x}_m)) \tag{3.6}$$

and reconstructing a surrogate function $g(\boldsymbol{x}_{m+1})$ using the point of tangency $f(\boldsymbol{x}_{m+1}) = g(\boldsymbol{x}_{m+1}|\boldsymbol{x}_{m+1})$.

However, a step in the right direction accomplishes the same goal. A Newton step along the surrogate function would move in the right direction

$$\boldsymbol{x}_{m+1} = \boldsymbol{x}_m - \frac{\frac{\partial g}{\partial \boldsymbol{x}_m}}{\frac{\partial^2 g}{\partial \boldsymbol{x}_m^2}}. \tag{3.7}$$

This new point $\boldsymbol{x}_{m+1}$ is sufficient for creating a surrogate function $g(\boldsymbol{x}_{m+1})$, using the point of tangency $f(\boldsymbol{x}_{m+1}) = g(\boldsymbol{x}_{m+1}|\boldsymbol{x}_{m+1})$.

Under the dual problem of minimization, the surrogate function we seek will be one that majorizes f. Majorization shares the structure of minorization. The function $h(\boldsymbol{x}|\boldsymbol{x}_m)$ majorizes $f(\boldsymbol{x})$ if it shares the value $f(\boldsymbol{x}_m)$ and for all other $\boldsymbol{x}$, h is above f. More precisely, we require

$$\begin{aligned} f(\boldsymbol{x}_m) &= h(\boldsymbol{x}_m|\boldsymbol{x}_m) \\ f(\boldsymbol{x}) &\leq h(\boldsymbol{x}|\boldsymbol{x}_m), \boldsymbol{x} \neq \boldsymbol{x}_m. \end{aligned} \tag{3.8}$$

Under the minimization formulation, we follow the same procedure as with maximization, driving the surrogate function downhill.

3.2.3 Exploring MM techniques

The marginal utility of the MM algorithm hinges on the selection of the surrogate function. There are many ways to find a majorizing or minorizing function for a given problem of interest; the key is selecting the surface to accomplish a particular goal. We will discuss two well-known MM tools, which we use together in our implementation. For a concave function, the simple tangent line satisfies the requirements of a majorizing function [Hunter and Lange, 2004]. Specifically, for a concave $f(\boldsymbol{x})$, the tangent line

$$g(\boldsymbol{x}|\boldsymbol{x}_m) = f(\boldsymbol{x}_m) + df(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) \tag{3.9}$$

satisfies the majorizing requirements. That is, we have $g(\boldsymbol{x}_m|\boldsymbol{x}_m) = f(\boldsymbol{x}_m)$ and $g(\boldsymbol{x}|\boldsymbol{x}_m) \geq f(\boldsymbol{x}) \, \forall \boldsymbol{x} \neq \boldsymbol{x}_m$. When using this approach, finding $argmin(g(\boldsymbol{x}|\boldsymbol{x}_m))$ is obviously problematic. This will not be a concern for us, as we will use this technique in conjunction with others.

The second method relies on Jensen's Inequality, which states that a secant through a convex function lies above the arc of the function bounded by the two points of intersection [Hunter and Lange, 2004]. For a convex function $f(t)$, this is

$$f(\sum_i \alpha_i t_i) \leq \sum_i \alpha_i f(t_i). \tag{3.10}$$

We take this inequality and transform it into a recipe for dealing with inner products within convex functions. First, we can transform $\boldsymbol{x}$ with a linear function $\boldsymbol{c}'\boldsymbol{x}$ for some $\boldsymbol{c}$. Setting $\alpha_i = c_i x_{m_i} / \boldsymbol{c}'\boldsymbol{x_m}$ gives the majorization

$$g(\boldsymbol{x}|\boldsymbol{x}_m) = \sum_i \alpha_i f\left(\frac{c_i}{\alpha_i}(x_i - x_{m_i}) + \boldsymbol{c}'\boldsymbol{x_m}\right). \tag{3.11}$$

Again, we can see that $g(\boldsymbol{x}|\boldsymbol{x}_m)$ fulfills the MM criteria: $g(\boldsymbol{x}|\boldsymbol{x}_m) = f(\boldsymbol{c}'\boldsymbol{x})$ and $g(\boldsymbol{x}|\boldsymbol{x}_m) \geq f(\boldsymbol{c}'\boldsymbol{x}) \leq \forall \boldsymbol{x} \neq \boldsymbol{x}_m$.

3.2.4 Developing an MM algorithm for the SCCS model

We consider both of these tricks in the context of the self-controlled case series likelihood. The goal when constructing our surrogate surface is decoupling the covariates. That is, we preserve the Newton step framework from CCD implementation, and we want the Newton steps along each coordinate to be locally independent. Looking at Equation (3.2), we see that the numerator contribution $\sum_{k=1}^{K_i}\left(y_{ik}\,\boldsymbol{x}'_{ik}\boldsymbol{\beta}\right)$ is already decoupled. Therefore, we focus on separating variables out of the log denominator term in Equation (3.2). That is, we want to apply MM transformations to the log denominator term so that each Newton step relies on the other coordinates only through the previous iteration's solutions. We do this by taking two MM transformations of the log likelihood denominator.

For clarity, we specify notation for the sums that appear the numerator and denominator contributions for each patient i. The sum that appears in the numerator is

$$T_i = \sum_{k=1}^{K_i}\left(y_{ik}\,\boldsymbol{x}'_{ik}\boldsymbol{\beta}\right), \tag{3.12}$$

and, the sum from the denominator is

$$D_i = \sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}. \tag{3.13}$$

Under this new notation, our log likelihood becomes

$$L(\boldsymbol{\beta}) = \sum_{n=1}^{N} T_i - n_i \log(D_i). \tag{3.14}$$

We begin constructing our minorizing surface to Equation (3.14) by addressing the outer logarithm. Using the tangent line inequality from Equation (3.9), we see that

$$-\log\left(D_i\right) \geq -\log\left(D_i^m\right) - \frac{1}{D_i^m}\left(D_i - D_i^m\right) \tag{3.15}$$

where

$$D_i^m = \sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}^m}. \tag{3.16}$$

Checking the minorizing constraint

$$L\left(\boldsymbol{\beta}\right) \geq \sum_{i=1}^{N}\left\{T_i + n_i\left[-\log\left(D_i^m\right) - \frac{1}{D_i^m}\left(\sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}} - D_i^m\right)\right]\right\} \tag{3.17}$$

we see that we have an appropriate minorization. We can confirm the tangential requirement

$$L\left(\boldsymbol{\beta}^m\right) = \sum_{i=1}^{N}\left\{T_i^m + n_i\left[-\log\left(D_i^m\right) - \frac{1}{D_i^m}\left(\sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}^m} - D_i^m\right)\right]\right\}. \tag{3.18}$$

Using the tangent line technique allows us to cope with the log term at each iteration, but the covariate updates would still be coupled. There remains an inner product within the exponential, and we need to decouple this term. We return to Equation (3.11), recognizing that e^x is a convex function. Again, we define notation to simplify the exposition. Let

$$S_{ik} = l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}} \tag{3.19}$$

so that

$$S_{ik}^m = l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}^{\boldsymbol{m}}}. \tag{3.20}$$

Note that

$$\sum_{k=1}^{K_i} S_{ik} = D_i. \tag{3.21}$$

With this notation and using the result of Equation (3.11), we see that

$$S_{ik} \leq \sum_{j=1}^{J} \alpha_j e^{\frac{x_{ikj}}{\alpha_j}\left(\beta_j - \beta_j^m\right)} S_{ik}^m \tag{3.22}$$

where

$$\alpha_{ikj} = \frac{|x_{ikj}|^p}{\sum_{h=1}^{J} |x_{ikh}|^p} \tag{3.23}$$

for some integer p. For indicator $\boldsymbol{X}$ the choice of p is inconsequential, and we have

$$\alpha_{ikj} = \frac{x_{ikj}}{z_{ik}} \tag{3.24}$$

for z_{ik} the count of drugs present in era k for patient i. Using the results of both MM techniques together, we have the full MM surface

$$\begin{aligned} Q(\beta|\beta^m) = & \sum_{i=1}^{N} \{T_i + n_i [- \log (D_i^m) \\ & - \frac{1}{D_i^m} \left(\sum_{k=1}^{K_i} \sum_{j=1}^{J} \alpha_{ikj} e^{\left(\frac{x_{ikj}}{\alpha_{ikj}}\left(\beta_j - \beta_j^m\right)\right)} S_{ik}^m - D_i^m \right) \Bigg] \Bigg\} \end{aligned} \tag{3.25}$$

or

$$\begin{aligned} Q(\beta|\beta^m) = & \sum_{i=1}^{N} \{T_i + n_i [- \log (D_i^m) \\ & - \frac{1}{D_i^m} \left(\sum_{k=1}^{K_i} \sum_{j=1}^{J} \frac{x_{ikj}}{z_{ik}} e^{\left(z_{ik}\left(\beta_j - \beta_j^m\right)\right)} S_{ik}^m - D_i^m \right) \Bigg] \Bigg\}. \end{aligned} \tag{3.26}$$

We see that $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^m)$ satisfies the two requirements. That is,

$$\begin{aligned} L(\boldsymbol{\beta}) &\geq Q(\boldsymbol{\beta}|\boldsymbol{\beta}^m), \text{and} \\ L(\boldsymbol{\beta}^m) &= Q(\boldsymbol{\beta}^m|\boldsymbol{\beta}^m). \end{aligned} \tag{3.27}$$

Most importantly, we see that the covariates are decoupled.

3.2.5 Newton steps in the MM approach

Within the MM framework, we keep the strategy of updating with Newton steps. Rather than maximize the surrogate function, we take a single Newton step along each covariate to advance our position. To obtain the Newton's steps, we require both the partial derivative and unidirectional Hessian for each covariate. The partial derivative of the surrogate function for covariate l is

$$\frac{\partial Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{\boldsymbol{m}})}{\partial \beta_l} = \sum_{i=1}^{N} \left\{ \sum_{k=1}^{K_i} y_{ik}\, x_{ikl} - \frac{n_i}{D_i^m} \sum_{k=1}^{K_i} \alpha_{ikj} \frac{x_{ikl}}{\alpha_{ikj}} e^{\left(\frac{x_{ikj}}{\alpha_{ikj}}\left(\beta_j - \beta_j^m\right)\right)} S_{ik}^m \right\} \tag{3.28}$$

or, equivalently,

$$\frac{\partial Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{\boldsymbol{m}})}{\partial \beta_l} = \sum_{i=1}^{N} \left\{ \sum_{k=1}^{K_i} y_{ik}\, x_{ikl} - \frac{n_i}{D_i^m} \sum_{k=1}^{K_i} x_{ikl} e^{\left(z_{ik}\left(\beta_j - \beta_j^m\right)\right)} S_{ik}^m. \right\} \tag{3.29}$$

When evaluated at our current location,

$$\left.\frac{\partial Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{\boldsymbol{m}})}{\partial \beta_l}\right|_{\beta_l{}^m} = \sum_{i=1}^{N} \left\{ \sum_{k=1}^{K_i} y_{ik}\, x_{ikl} - \frac{n_i}{D_i^m} \sum_{k=1}^{K_i} x_{ikl} S_{ik}^m \right\}. \tag{3.30}$$

Similarly, the local curvature is

$$\frac{\partial^2 Q(\boldsymbol{\beta}|\beta^m)}{\partial \beta_l{}^2} = \sum_{i=1}^{N} -\frac{n_i}{D_i^m} \left\{ \sum_{k=1}^{K_i} z_{ik} x_{ikl} e^{\left(z_{ik}\left(\beta_j - \beta_j^m\right)\right)} S_{ik}^m \right\}. \tag{3.31}$$

Evaluating this curvature at our current location gives

$$\left.\frac{\partial^2 Q(\boldsymbol{\beta}|\beta^m)}{\partial \beta_l{}^2}\right|_{\beta_l{}^m} = \sum_{i=1}^{N} -\frac{n_i}{D_i^m} \left\{ \sum_{k=1}^{K_i} z_{ik} x_{ikl} S_{ik}^m \right\}. \tag{3.32}$$

The regularization component of the Newton step remains the unchanged from the CCD implementation, so we do not show it here.

As we see in Equations (3.30, 3.32), the gradient and Hessian components of our Newton steps for a particular β_l rely only on β_j $j \neq l$ through the constants β^m. Thus, we have succeeded in our main goal for employing the

MM algorithm: decoupling our updates to the covariates. Our change to β_l^m is now

$$\Delta^m \beta_l^m = - \left. \frac{\frac{\partial}{\partial \beta_l} \left(Q(\boldsymbol{\beta}|\beta^m) + \log[p(\boldsymbol{\beta})] \right)}{\frac{\partial^2}{\partial {\beta_l}^2} \left(Q(\boldsymbol{\beta}|\beta^m) + \log[p(\boldsymbol{\beta})] \right)} \right|_{{\beta_l}^m} . \tag{3.33}$$

We modify the fitting procedure from Suchard et al. [2013] shown in Algorithm 1 in our new MM algorithm shown in Algorithm 2. Again, we declare convergence when the sum of the absolute change in $\boldsymbol{X\beta}$ from successive iterations falls below 1×10^{-6}.

Algorithm 2 MM algorithm for fitting the SCCS model. This highlights the parallel steps of the algorithm. In particular, we see the three parallelized targets for computation.

```
Initialize: $m = 0$
Initialize: $\boldsymbol{\beta^m} = 0$
while not $\boldsymbol{X\beta}$ converged do
    In parallel compute $\frac{\partial}{\partial \beta_j} \left( Q(\boldsymbol{\beta}|\beta^m) + \log[p(\boldsymbol{\beta})] \right)$ and $\frac{\partial^2}{\partial {\beta_j}^2} ( Q(\boldsymbol{\beta}|\beta^m) + \log[p(\boldsymbol{\beta})])$ and update $\beta_j^{m+1} = \beta_j^m + \Delta^m \beta_j^m$
    In parallel compute $\boldsymbol{X\beta}$
    Atomically add $\sum_{k=1}^{K_i} l_{ik} e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}} \; \forall i$
    $m = m + 1$
end while
```

Multi-core implementation

Now that we have decoupled the Newton steps, we can move away from the serial iterative approach of CCD toward a parallel iterative implementation. We expose three targets for parallelization.

3.2.5.1 Decoupled update

First, each Newton step or augmented Newton step can proceed in parallel. Each update is independent between covariates. Therefore, we can assign each update as a separate tasks to computing elements. The work per thread becomes

$$\beta_l^{m+1} = \beta_l^m - \frac{\sum_{i=1}^{N} \left\{ \sum_{k=1}^{K_i} y_{ik}\, x_{ikl} - \frac{n_i}{D_i^m} \sum_{k=1}^{K_i} x_{ikl} S_{ik}^m \right\} + \frac{\partial}{\partial \beta_l} (\log[p(\boldsymbol{\beta})])}{\sum_{i=1}^{N} - \frac{n_i}{D_i^m} \left\{ \sum_{k=1}^{K_i} z_{ik} x_{ikl} S_{ik}^m \right\} + \frac{\partial^2}{\partial \beta_l^2} (\log[p(\boldsymbol{\beta})])}. \tag{3.34}$$

3.2.5.2 Matrix-vector multiplication

The CCD approach efficiently updates $\boldsymbol{X\beta}$ by incrementing with each partial step. After a Newton step for covariate j,

$$\boldsymbol{X\beta} = \boldsymbol{X\beta} + \Delta\beta_j \boldsymbol{X_j}. \tag{3.35}$$

In this setting, we are not required to recompute $\boldsymbol{X\beta}$. However, this advantage breaks down when using decoupled updates. We may simultaneously update a group of covariates in $\boldsymbol{\beta}$. Waiting after these steps to iteratively update $\boldsymbol{X\beta}$ would negate the benefits of working in parallel.

Therefore, we no longer update $\boldsymbol{X\beta}$ with each step, but rather recompute after updating $\boldsymbol{\beta}$. This matrix-vector multiplication is amenable to parallelization. We allocate each $\boldsymbol{X}_k\boldsymbol{\beta}$ as the parallelized task, with each $\boldsymbol{X}_k\boldsymbol{\beta}$ handled independently.

3.2.5.3 Sum of exponentials

We can also compute

$$D_i^m = \sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}_{ik}'\boldsymbol{\beta}^m} \tag{3.36}$$

in parallel for each $i \in \{1, \ldots, N\}$. However, there is a notable caveat. For updates to D_i^m by k, there is a race condition, where updates from separate eras for the same patient try to access the same partial sum. To account for this, we atomically add $l_{ik}\, e^{\boldsymbol{x}_{ik}'\boldsymbol{\beta}^m}$ for every k to D_i^m. These sums D_i^m are independent between different i, allowing the majority of these updates to occur simultaneously.

MM with acceleration

A well-known side effect of the MM algorithm is bloating of the iteration count [Lange, 1995]. Each Newton step is conservative; we could move further up the surface than the surrogate function permits. The result is slow convergence. Many have tried to address this problem by developing acceleration techniques, tricks that allow the MM step to be larger.

One straightforward approach to accelerating an MM algorithm is to double the size of the MM step. That is, instead of updating

$$\beta_l^{m+1} = \beta_l^m + \Delta^m \beta_j^m \tag{3.37}$$

we update

$$\beta_l^{m+1} = \beta_l^m + 2 * \Delta^m \beta_j^m, \tag{3.38}$$

following Lange [1995] and Lange and Wu [2008]. This approach nullifies the MM guarantee of ascent. For some problems, experience dictates that this violation is not practically significant [Lange and Wu, 2008]. When this is used absent regularization, doubling the MM step is equivalent to halving the Hessian. We take this idea and apply it to the regularized setting, but we would prefer to preserve the guarantees of the MM algorithm.

We extend this approach by dividing each of our unidirectional MM Newton Hessian values by a factor $\phi \geq 1$. Now, we update

$$\beta_l^{m+1} = \beta_l^m - \frac{\frac{\partial}{\partial \beta_l}\left(Q(\boldsymbol{\beta}|\boldsymbol{\beta}^m) + \log[p(\boldsymbol{\beta})]\right)}{\frac{\partial^2}{\partial {\beta_l}^2}\left(\frac{1}{\phi}Q(\boldsymbol{\beta}|\boldsymbol{\beta}^m) + \log[p(\boldsymbol{\beta})]\right)}. \tag{3.39}$$

After completing MM Newton steps for each covariate, we recompute the log posterior density. If $L(\boldsymbol{\beta}^m) > L(\boldsymbol{\beta}^{m+1})$, we recognize that the augmented Newton step was too ambitious. We could discard or accept this step, but we choose to accept it here. This moves our position to the other side of the zenith. However, we reset $\phi = \frac{\phi}{2}$, ensuring that our next step is more conservative. We always decrease the augmentation factor. At worst, this pushes ϕ to 1. This ensures that we drive the optimization uphill, even if we take a small number of inappropriate steps. Since we already compute many components of the log likelihood after a full cycle through $\boldsymbol{\beta}$, the marginal cost to check for uphill movement is minimal.

Algorithm 3 MM algorithm for fitting the SCCS model, similar to the one shown in Algorithm 2. Here we include updating the augmentation factor ϕ.

```
Initialize: $m = 0$
Initialize: $\phi = \phi_0$
Initialize: $\boldsymbol{\beta}^{\boldsymbol{m}} = 0$
while not $\boldsymbol{X}\boldsymbol{\beta}$ converged do
    In parallel compute $\frac{\partial}{\partial \beta_l}(Q(\boldsymbol{\beta}|\beta^m) + \log[p(\boldsymbol{\beta})])$ and $\frac{\partial^2}{\partial {\beta_l}^2}(\frac{1}{\phi}Q(\boldsymbol{\beta}|\beta^m) + \log[p(\boldsymbol{\beta})])$ and update $\beta_j^{m+1} = \beta_j^m + \Delta^m \beta_j^m$
    In parallel compute $\boldsymbol{X}\boldsymbol{\beta}$
    Atomically add $\sum_{k=1}^{K_i} l_{ik}\, e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}\ \forall i$
    $m = m + 1$
    if $L(\boldsymbol{\beta}^m) + \log[p(\boldsymbol{\beta}^m)] > L(\boldsymbol{\beta}^{m+1}) + \log[p(\boldsymbol{\beta}^{m+1})]$ then
        $\phi = \frac{\phi}{2}$
    end if
end while
```

MM with quasi-Newton acceleration

A class of techniques for acceleration that also preserve the ascent guarantee are quasi-Newton methods, and these are useful for accelerating MM algorithms [Zhou et al., 2011]. The unifying idea of quasi-Newton methods is the use of secant approximations. We develop our quasi-Newton acceleration closely following Zhou et al. [2011].

We begin by considering Newton's method in the context of root finding for $\mathbf{0} = \boldsymbol{\beta} - F(\boldsymbol{\beta})$. Following Zhou et al. [2011], $F(\boldsymbol{\beta})$ is an algorithm map. In the context of the SCCS model, $F(\boldsymbol{\beta})$ is the resulting position from a cycle of uncoupled Newton steps along the MM surface tangent at $\boldsymbol{\beta}$. Newton's method solutions for root finding proceed as

$$\boldsymbol{\beta}^{m+1} = \boldsymbol{\beta}^m - [\boldsymbol{I} - dF(\boldsymbol{\beta}^m)]^{-1}[\boldsymbol{\beta}^m - F(\boldsymbol{\beta}^m)],$$

where the Hessian matrix $dF(\boldsymbol{\beta}^m)$ may be challenging to invert. Quasi-Newton methods avoid inverting $dF(\boldsymbol{\beta}^m)$ by substituting a low rank secant approximation $\boldsymbol{M}$. With $\boldsymbol{M}$ in hand, we could substitute $(\boldsymbol{I} - \boldsymbol{M})$ for $(\boldsymbol{I} - dF(\boldsymbol{\beta}^m))$ and compute the easier $(\boldsymbol{I} - \boldsymbol{M})^{-1}$.

We consider $\boldsymbol{U} = (\boldsymbol{u}_{m-q}, \ldots, \boldsymbol{u}_m)$ and $\boldsymbol{V} = (\boldsymbol{v}_{m-q}, \ldots, \boldsymbol{v}_m)$, matrices of the q most recent $\boldsymbol{u}$ and $\boldsymbol{v}$ vectors where $\boldsymbol{u}_m = F(\boldsymbol{\beta}^m) - \boldsymbol{\beta}^m$ and $\boldsymbol{v}_m = F(F(\boldsymbol{\beta}^m)) - F(\boldsymbol{\beta}^m)$. Here, the secant requirements are $\boldsymbol{M}\boldsymbol{u}_m = \boldsymbol{v}_m$. Relying on Proposition 1 from Zhou et al. [2011], we observe that $\boldsymbol{M} = \boldsymbol{V}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'$ provides our secant approximation to $dF(\boldsymbol{\beta}^m)$. Our quasi-Newton update is

$$\boldsymbol{\beta}^{m+1} = \boldsymbol{\beta}^m - [\boldsymbol{I} - \boldsymbol{V}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}']^{-1}[\boldsymbol{\beta}^m - F(\boldsymbol{\beta}^m)].$$

Finding the optimal q is not obvious. For different problems, the q that produces the fewest iterations to convergence may vary [Zhou et al., 2011].

3.3 Demonstration

3.3.1 Comparing steps graphically

We begin with a tiny synthetic study to compare the MM steps graphically. With a microscopic dataset of two patients and two drug exposures, we can easily visualize the log posterior. On this surface, we plot the steps that our optimization methods take to attain the maximum in Figure (3.1). Specifically, we plot the CCD steps and the MM approach with no augmentation factor ($\phi = 1$). Both algorithms start at the origin, and each point represents the position after a full traversal through $\boldsymbol{\beta}$. We specify the prior variance as $\lambda = 1$. For clarity in the graphic, we limit the iterations to 4. At 4 iterations, the CCD method converges to the true optimum. However, the MM approach without acceleration does not.

3.3.2 Diclofenac and gastrointestinal bleeding

Non-steroidal anti-inflammatory drugs (NSAIDs) are a cornerstone of pain management, providing both anti-inflammatory and analgesic action. They are commonly used in management of osteoarthritis and rheumatoid arthritis,

Algorithm 4 MM algorithm for fitting the Bayesian self-controlled case series model using the quasi-Newton approximation.

```
Initialize: $m = 0$
Initialize: $\boldsymbol{\beta}^m = \mathbf{0}$
Update $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 + \Delta^0 \boldsymbol{\beta}^0$
Compute $\boldsymbol{u} = \boldsymbol{\beta}^1 - \mathbf{0}$ and add to set $\boldsymbol{U}$
**for** $p \in (2, \ldots, q-1)$ **do**
    Update $\boldsymbol{\beta}^p = \boldsymbol{\beta}^{p-1} + \Delta^{p-1} \boldsymbol{\beta}^{p-1}$
    Compute $\boldsymbol{u} = \boldsymbol{\beta}^p - \boldsymbol{\beta}^{p-1}$ and add to set $\boldsymbol{U}$
    Compute $\boldsymbol{v} = \boldsymbol{u}$ and add to set $\boldsymbol{V}$
**end for**
Update $\boldsymbol{\beta}^q = \boldsymbol{\beta}^{q-1} + \Delta^{q-1} \boldsymbol{\beta}^{q-1}$
Compute $\boldsymbol{v} = \boldsymbol{\beta}^q - \boldsymbol{\beta}^{q-1}$ and add to set $\boldsymbol{V}$
**while** $\boldsymbol{X}\boldsymbol{\beta}$ not converged **do**
    Update $\boldsymbol{\beta}^m = \boldsymbol{\beta}^{m-1} + \Delta^{m-1} \boldsymbol{\beta}^{m-1}$
    Compute $\boldsymbol{u} = \boldsymbol{\beta}^m - \boldsymbol{\beta}^{m-1}$ and add to set $\boldsymbol{U}$
    Update $\boldsymbol{\beta}^{m+1,(MM)} = \boldsymbol{\beta}^m + \Delta^m \boldsymbol{\beta}^m$
    Compute $\boldsymbol{v} = \boldsymbol{\beta}^{m+1} - \boldsymbol{\beta}^m$ and add to set $\boldsymbol{V}$
    Compute $\boldsymbol{M} = \boldsymbol{U}'(\boldsymbol{U} - \boldsymbol{V})$
    Compute $\boldsymbol{\beta}^{m+1,(QN)} = \boldsymbol{\beta}^m + \boldsymbol{V}\boldsymbol{M}^{-1}\boldsymbol{U}'(\boldsymbol{\beta}^m - \boldsymbol{\beta} m - 1)$
    **if** $L(\boldsymbol{\beta}^{m+1,(MM)}) + \log[p(\boldsymbol{\beta}^{m+1,(MM)})] > L(\boldsymbol{\beta}^{m+1,(QN)}) + \log[p(\boldsymbol{\beta}^{m+1,(QN)})]$ **then**
        Update $\boldsymbol{\beta}^{m+1} = \boldsymbol{\beta}^{m+1,(MM)}$
    **else**
        Update $\boldsymbol{\beta}^{m+1} = \boldsymbol{\beta}^{m+1,(QN)}$
    **end if**
**end while**
```

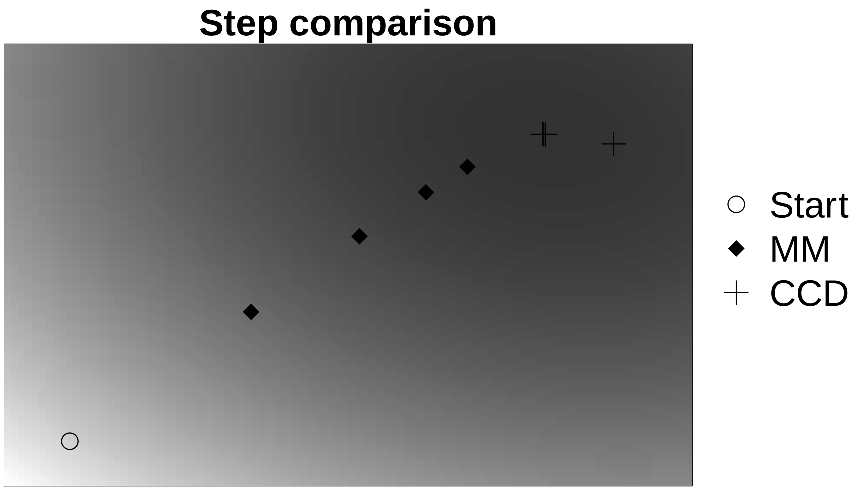


Figure 3.1: *Step comparison between the Cyclic Coordinate Descent (CCD) and minorization-maximization (MM) without acceleration ($\phi = 1$) using our toy two dimensional dataset. For each method, we limit the iterations to 4, showing how far each progresses in the same number of iterations. The CCD approach converges after 4 iterations; the MM approach does not. The common starting point for both approaches is (0,0), shown with an open circle, and each point represents the position after a full traversal through $\boldsymbol{\beta}$.*

among other conditions [Hawkey et al., 1998]. This class of medications includes naproxen and ibuprofen, common over-the-counter medications, as well as diclofenac, a more potent variety available with prescription. Among the most common adverse events associated with these drugs are events related to the gastrointestinal (GI) system [Hawkey et al., 1998, Lanas, 2010]. Minor adverse events from these medications include nausea and abdominal pain. However, more serious adverse events can follow from NSAID use as well. GI bleeding is among the most common serious adverse events associated with NSAIDs.

We contribute to the risk estimates of diclofenac for GI bleeding by considering it in the context of our longitudinal healthcare datasets. We want to compare our MM approaches with the CCD method fitting the SCCS model on GI bleeding events among users of diclofenac. To perform these studies, we examine the MarketScan Lab Results (MSLR) dataset, maintained by the Reagan-Udall Foundation Innovation in Medical Evidence Development and Surveillance project. This database contains time-stamped patient data, including laboratory results, drug exposures, and diagnoses, deidentified to compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA), comprising 1.5 million patient lives. The MSLR dataset includes both inpatient and outpatient records.

The development of a common data model (CDM) through the Observational Medical Outcomes Partnership (OMOP) experiment facilitates statistical methods implementation using these data [Stang et al., 2010]. The CDM allows us to address pertinent questions about health practices, including comparative drug safety, by standardizing data concept representation across resources [Overhage et al., 2012]. Standardization takes the native representation in a clinical data set of concepts like medication ingredients and diagnosis definitions, such as International Classification of Diseases version 9 (ICD-9), and translates them to a common representation. This facilitates consistent and reproducible analysis across datasets; we can apply the same analyses to different data resources, without having to recode our approach to accommodate dataset-specific variations. We depend on the OMOP CDM version 4 for representation of concepts of interest.

One aspect of the OMOP CDM version 4 requires special attention. We use the OMOP CDM 4 definition of a drug era. A drug era is a combination of individual prescriptions or drug fills. For example, if the same medication is refilled routinely at the end of its 30 day supply for 2 refills, this appears as a single 90 day drug era. OMOP uses a standard 30 day persistence window, where if a new supply of the same medication is given within 30 days of the termination of a previous supply, it is considered the same era. For example, consider a patient who takes metformin for 60 days, forgets to refill a prescription for 4 days and does not take any medication. Then on the 5th day, that patient refills the prescription and continues taking metformin for 90 days. With a 30 day persistence window, all of the medication use actions result in a single 154 day drug era. The 30 day persistence window helps buffer refill discontinuities.

We select all patients who experienced a diagnosis that the OMOP common data model version 4 considered a subset of GI bleeding and who were exposed to diclofenac. There are $N = 120,034$ such patients. To control for the other medications that may contribute to risk for GI bleeding, we include all other drug exposures. In the MSLR dataset, this population used $J = 940$ distinct drugs. Over all of these patients, there are $K = 5,681,213$ distinct drug eras.

We use cross-validation based on the predictive log-likelihood of the hold-out set to select the prior variance $\frac{1}{\tau}$. Suchard et al. [2013] use a log-scale grid search that is computationally expensive even with only a single parameter. To help overcome this burden, we turn to Genkin et al. [2007] in implementing an “autosearch" for hyperparameter selection. We start with an initial guess and then increase or decrease our guess by one log unit until we have bracketed the maximum of the hold-out set predicted log-likelihood. Then we compute a quadratic approximation to the predicted log-likelihood. The maximum of this approximate surface becomes our estimate. For this problem, our estimated $\frac{1}{\tau}$ is 1.21. We only perform this cross-validation once using the CCD fitting approach, and this computational cost does not enter into our calculations. For all following performance comparisons, we pre-specify this prior variance. We find the point estimate of the log relative risk for Diclofenac in this population is 0.17 with a bootstrap 95 % confidence interval of $[0.02, 0.26]$.

3.3.2.1 Parallelization

To take advantage of all of our parallelization opportunities, we turn to Amazon Web Services. For all of our timing comparisons, we elect to use 36 core, compute-optimized, Intel Xeon E5-2666v3, 60 GiB memory, EBS-only instances. We test each of the parallelization tasks using 1, 2, 4, 8, 16, and 32 threads. To measure performance gain from parallelization alone, our baseline speed is the MM algorithm with $\phi = 1$ using no parallelization. The speedup as a function of parallelization appears in Figure (3.2).

The speedup from parallelization alone in the context of the MM algorithm follows a similar pattern for both the decoupled Newton steps and the $\boldsymbol{X\beta}$ computation. For both of these tasks, the relative gain in speed plateaus around 6 fold using 16 cores. Of the full fitting time in the non-parallelized (single core) implementation, 83% of time is spent in either the decoupled Newton steps or the $\boldsymbol{X\beta}$ computation, with only 15% of time is spent summing the exponentials. Therefore, it is reasonable that the overall speedup of the MM algorithm from parallelization tracks with the speedup from the decoupling

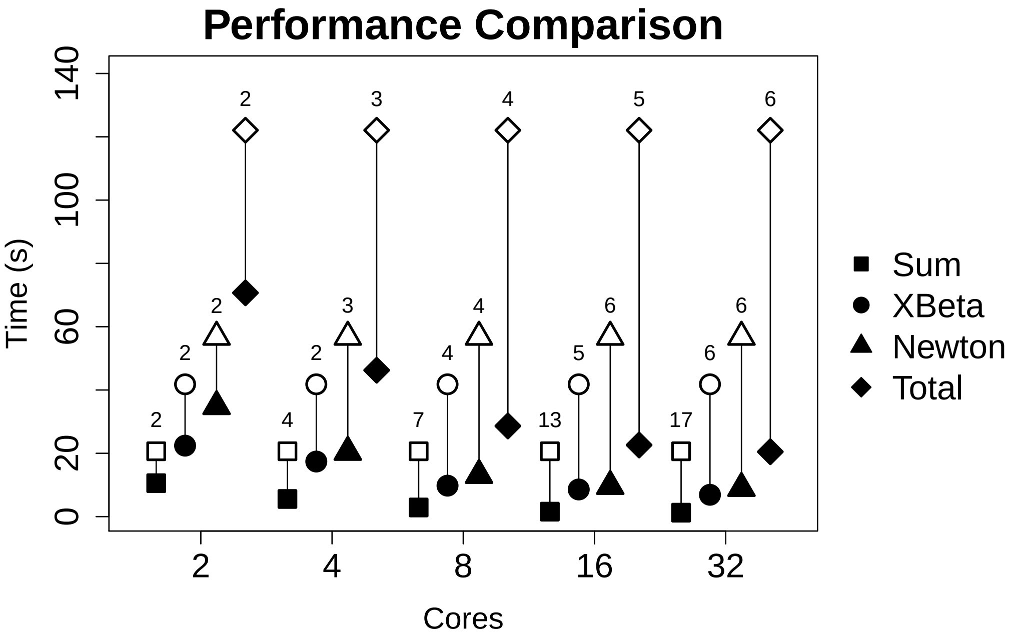


Figure 3.2: *Convergence time and fold speedup for each parallelization task as well as total mode-finding time for the MM implementation without acceleration over different thread counts. The open symbols represent the single core times, and the closed symbols represent the times at each core count shown on the axis. The fold speed up relative to the single core times is above the symbols. The speedup for the decoupled Newton steps and the $\boldsymbol{X\beta}$ computation both plateau at just above 6 fold speed-up around 16 threads. Since these dominate the computational load, the total time follows a similar pattern. The atomic addition of exponentials outperforms these calculations with a maximum speedup of over 17 fold at 32 cores.*

and the $\boldsymbol{X}\boldsymbol{\beta}$ computation.

It is somewhat surprising that the atomic summing of exponentials significantly outperforms the decoupled Newton steps and the $\boldsymbol{X}\boldsymbol{\beta}$ computation. This process showed in excess of 17 fold speedup using 32 threads. Using the full 36 cores did not improve the performance. In fact, using 36 cores was slightly slower than using 32 cores. This is helpful when looking at Figure (3.2), where it appears that the performance improvement is strictly increasing with the number of cores used.

3.3.2.2 Acceleration

We first turn to the performance of the augmented Newton step acceleration. Considering $\phi \in \{1, 2, 4, 8, 16, 32\}$, we see that the time to convergence is not monotonic in ϕ. For $\phi = 1$, we have the non-accelerated MM algorithm. Increasing to $\phi = 2$ slows down the convergence rate. However, between 4 and 16, the number of iterations required for convergence decreases by roughly a factor of three. The greatest improvement over the MM algorithm appears at $\phi = 4$, where 51 iterations produce convergence.

Similarly, we can look at the performance of the quasi-Newton acceleration, and we see that the number of iterations is not perfectly related to the order of Newton approximation q. The greatest improvement over the non-quasi-Newton MM algorithm appears at $q = 2$, where only 32 iterations produced convergence. Beyond $q = 2$, the effect of the quasi-Newton approach fluctuates.

3.3.2.3 Time

Finally, we compare the raw fitting time of the CCD approach with the MM algorithm, considering the non-accelerated version as well as the two acceleration methods. For the acceleration methods, we cherry-pick the best performing settings as shown above. This helps capture the best-case scenario for speedup.

The CCD fitting time is the standard, requiring 83 iterations and 195 seconds. We report the performance of the other methods as fold speed-up relative to this time. The simple MM algorithm produces up to 10 fold speedup, despite needing an additional 65 iterations. The quasi-Newton acceleration with $q = 2$

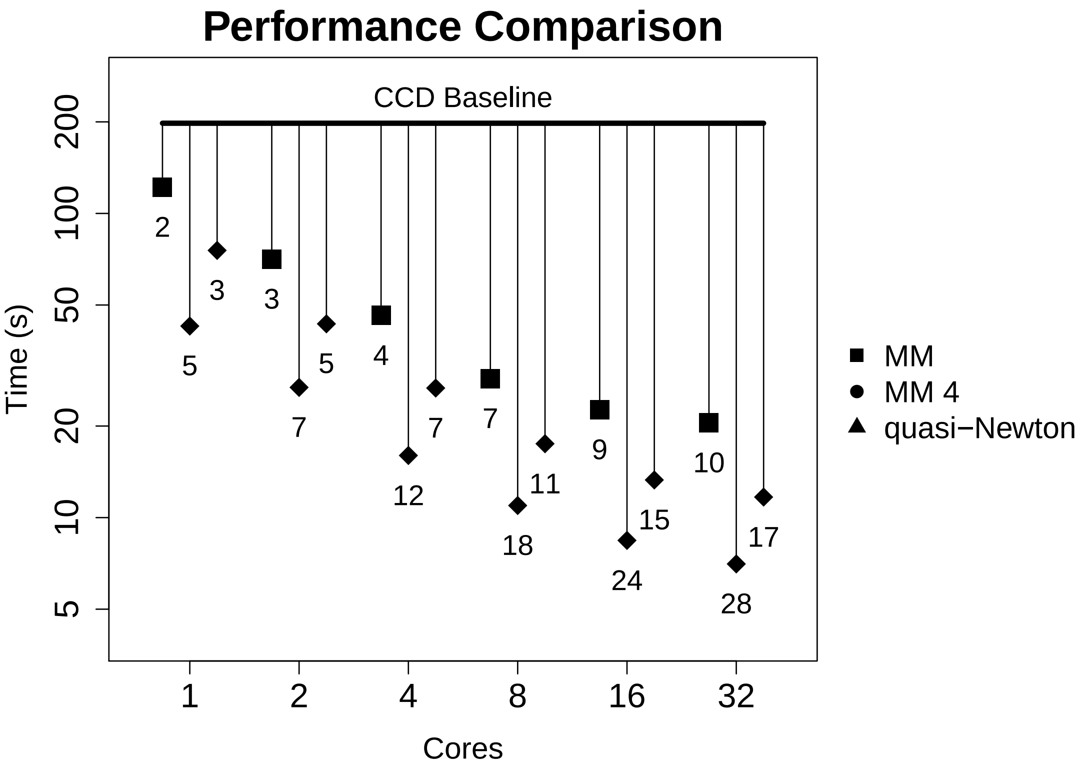


Figure 3.3: *We compare the fitting time for the MM with $\phi = 1$, MM with $\phi = 4$, and MM with quasi-Newton acceleration with $q = 2$ fitting procedures with parallelization (shown as symbols) to the CCD fitting time, shown as the horizontal line. We report the performance as fold speedup, shown in the printed values beneath the symbols versus the CCD algorithm.*

boosts the performance to 17 fold speedup versus the CCD fitting time, needing just 32 iterations for convergence. The augmented Newton steps with $\phi = 4$ produces the best speed-up, improving model fit time more than 28 fold with 51 iterations.

3.4 Discussion

In our two dimensional toy example, we see the challenge that the MM algorithm presents. Each of the MM Newton steps with $\phi = 1$ is excruciatingly more conservative than the CCD steps, which visually converges to the optimal point within two full updates. This captures how the MM Newton step and the CCD Newton steps differ. This also allows us to see how the acceleration choices discussed here ameliorate this situation. In particular, guessing $\phi = 4$ or using quasi-Newton with $q = 1$ in this problem produces updates that qualitatively replicate those of the CCD Newton steps. This helps explain how these correct for the conservative MM Newton steps.

Comparing the plain MM algorithm with the CCD approach, we see a notable decrease in cost per iteration, even without parallelization. Using CCD, the cost per iteration is roughly 2.3 seconds. For the MM algorithm, this drops to 0.8 seconds per iteration. This difference gives us an estimate of how much iteration bloating in the MM approach we can tolerate. Using these values, we break even at roughly 3 fold more iterations from the MM algorithm.

Recognizing when the MM approach produces faster results than CCD, or, conversely, when it slows model fitting, is critical. For some problems, the MM surrogate function may force our updates to be considerably more conservative than CCD. By identifying the computational cost break-even point, we can pivot from one approach to the other. This is particularly relevant for dynamically determining the procedure to use as a function of the local hardware. If the iteration count for the MM approach mildly exceeds the break-even point on a 32 core machine, the MM approach will likely be faster when leveraging all the parallelization opportunities. On the other hand, the same situation presented on a dual core laptop may make CCD the best choice. These comparisons are likely unreasonable for single model fitting events. However, if we are fitting multiple models serially, within a cross-validation, bootstrapping, or Markov chain Monte Carlo framework, it might be reasonable to adjust the fitting algorithm on the fly.

Considering the augmented Newton step acceleration approach, the fact that we guess $\phi = 4$ well in toy problem raises more questions than answers. One of the drawbacks of our approach is finding the initial step augmentation value ϕ. Guessing a poor value may lead to considerably longer time to convergence. Ideally, we would like to have a better method for selecting ϕ. Searching through the space of ϕ is not reasonable for a single fitting of the model. However, this becomes more appealing in a setting where a model may be fit multiple times. For example, we could double the initial ϕ at each model fitting, starting with $\phi = 1$. The efficiency gain from using a good ϕ might offset the inefficiency of searching for appropriate ϕ as the number of serial uses increased. Alternatively, we see that $\phi = 4$ helps the initial MM Newton step match with the size of the initial CCD step. We could extend this observation. If we take one full CCD step as well as one MM Newton step with $\phi = 1$, we could estimate ϕ from the length of the CCD update and the length of the MM update. With this as our initial guess, we can proceed with the step halving as appropriate if we overshoot the zenith.

Turning to the quasi-Newton acceleration, the efficiency fluctuation around q is nominally surprising. Intuitively, the higher q, the better the secant approximation should be. In practice, this is known to not be so simple [Zhou et al., 2011]. As with finding ϕ, selecting the best q is not obvious. Zhou et al. [2011] offer little guidance on how to find the optimal q, suggesting that it is problem-specific. Reflecting our experience with selecting ϕ, it is possible to imagine learning the best q as part of a sequence of model fittings.

The quasi-Newton acceleration with the best q requires the fewest iterations for convergence among all the approaches shown hear. While this reflects well on this approach, we should be cautious when comparing this count to the others. For each quasi-Newton step, we are effectively cycling through $\boldsymbol{\beta}$ twice. Additionally, we take q cycles through $\boldsymbol{\beta}$ before beginning the quasi-Newton steps. Therefore, the quasi-Newton iterations substantially underestimate the amount of work done to achieve an update if they are directly compared to the augmented step acceleration.

This additional work helps explain the performance discrepancy between the best quasi-Newton formulation and the best augmented Newton step formulation in Figure (3.3). The other work that differs between the two approaches includes the matrix inversion and several matrix-vector multiplications. Since we have constructed M to be low rank, its inverse is less costly to produce. However, this cost would certainly manifest for larger values of q.

Examining our regression risk estimate for diclofenac and GI bleeding, the log relative risk that we recover is lower than some reported values. Reviews of observational studies suggest the true relative risk diclofenac on upper GI bleeding is near 3.98 with 95% confidence interval $[3.36 - 4.72]$, while our recovered relative risk is 1.19 with 95% confidence interval $[1.02 - 1.30]$[Gonzalez et al., 2010]. While the fact that our method finds a significant positive association between diclofenac and GI bleeding is encouraging, the risk estimates are measurably different. There are several possible explanations for the discrepancy. First, this is a penalized estimate, biasing our value toward 0, which is consistent with underestimating the reported risk. Another possible result is that significant difference in controlling for other exposures exists between this study and those reported. Other research may not have controlled for as many exposures as we have. Finally, observational datasets frequently represent very different populations, with different group risk estimates. It is important to note the challenge observational datasets poses. Unmeasured confounding persistently complicates these results, our method may be unable to remove all unmeasured confounding. We use one major observational data modalities: claims data. These data are not collected for the scientific purposes. The claims data reflect billing practices and are limited by insurance acceptance policies. It is plausible that the source of bias from such collection is considerable.

In summary, the MM algorithm provides an elegant framework for developing optimization algorithms. However, integration of the MM algorithm into the tools for observational healthcare analysis has been slow. The primary concern is slow convergence in high-dimensional applications. Nevertheless, the potential for this algorithm remains, and some have pushed through the challenges in other disciplines [Zhou and Zhang, 2012, Zhou et al., 2010]. Our results highlight how an MM algorithm can improve speed fitting the SCCS model at scale, shedding light on a new resource for computing in the setting of massive observational data.

CHAPTER 4

Diabetes Treatment Trajectories

Well-established guidelines anchor clinical treatment for diabetes mellitus type II. However, characterizing actual treatment in practice is more elusive. To our knowledge, little research illustrates how patients progress through drug treatment regimens and how clinical tests alter these trajectories. We take a step forward by examining insurance claims and electronic health records data from four national datasets spanning 1994 to 2014, including a Medicare subset, private insurance claims, and data from General Electric Centricity users. This population consists of over 1 million patients who received at least one diagnosis of T2D at some time in their medical history and have both two hemoglobin A1c (HbA1c) tests and an oral anti-hyperglycemic mediation in their observed history prior to any insulin initiation. Across patients, we extract HbA1c measurements, the key justification for treatment intensification. Following the framework of guideline recommendations, we track the number of oral anti-hyperglycemic medications taken concurrently over the course of each patient‘s history. We model this treatment count over time jointly with most recent HbA1c status as a birth-death process, with first insulin use of any formulation as a terminating state. We track the percent at 0, 1, 2, 3 or 4+ of drugs over time and count the patients who transition from high to low HbA1c categories. We compute the transition rates between paired drug count and HbA1c state, which represent treatment intensification and de-intensification or response to treatment. We stratify by high/low persistent HbA1c status and compare intensification rates. The relative proportion of patients on 1,2,3, or 4+ drugs after a year on treatment stabilizes consistently across datasets. Between 49% and 64% of patients remain in their initial HbA1c state, independent of treatment. Furthermore, between 40% and 60% of patients who start with high HbA1c never achieve low HbA1c, even transiently. Patients with perennially high HbA1c often show both faster intensification and de-intensification. Therefore, we see that treatment impact on glycemic control may be moderate; diabetes patients fail to intensify treatment along previously recommended timelines; and poorly controlled patients struggle finding a consistent treatment strategy.

4.1 Introduction

Clear treatment guidelines from both the American Diabetes Association (ADA) and the American Association of Clinical Endocrinologists (AACE) direct management of diabetes mellitus type 2 (T2D) [Handelsman et al., 2015, American Diabetes Association, 2015]. For a treatment naive patient, the initial approach is monotherapy, with metformin usually recommended. Failure to meet hemoglobin A1c (HbA1c) goals triggers introduction of a second or third treatment at three month intervals. Persistent failure to attain HbA1c goals leads to insulin initiation.

Identifying how treatment intensification evolves in practice is more elusive [Grimes et al., 2015]. Current approaches have been largely limited to cross-sectional snapshots through longitudinal studies, surveys, and patient tracking databases [Turner et al., 1999, Alexander et al., 2008, Hampp et al., 2014, Dailey et al., 2002, Hazel-Fernandez et al., 2015]. The UK Prospective Diabetes Study (UKPDS) followed patients every three months for three, six, and and nine years post study initialization. From these patients, the UKPDS study demonstrates that 50% of patients need an additional treatment to manage their diabetes at three years, while 75% require an additional treatment at nine years [Turner et al., 1999]. However, this was a structured, randomized intervention, rather than an observation of real-world clinical data. To assay the clinical experience, Alexander et al. [2008] examine the National Disease and Therapeutic Index, a survey of office-based physicians selected by the American Medical Association and the American Osteopathic Association where the physicians report treatment patterns over two randomly sampled consecutive days per quarter. The authors compare these mixed samples from 1994 to 2007. They find that the proportion of patients on monotherapy declined from 82% in 1994 to 47% in 2007.

Massive observational resources, including insurance claims and electronic health records (EHRs) offer a new resource for understanding trajectories. Observational datasets have become fertile resources for addressing questions about diabetes in practice [Hampp et al., 2014, Grabner et al., 2013, Boccuzzi et al., 2001, Dailey et al., 2002, Pladevall et al., 2004, Maclean et al., 2004, Grimes et al., 2015, Slabaugh et al., 2015, Hazel-Fernandez et al., 2015, Weng et al., 2016]. Hampp et al. [2014] examine IMS Health Vector One National and Total Patient Tracker databases for annual prescription use from 2003 to 2012. Like Alexander et al. [2008], they are able to paint the trend in treatment using broad strokes. They identify a rising use of non-insulin anti-hyperglycemics and a surprisingly low rate of concomitant use of meformin with other treatment. Using eleven data sources including claims and electronic health records (EHR) data with a total of 250 million patients, Hripcsak et al. designed and implemented an observational study to map out treatment trajectories without information about how long a patient remains on a drug that the authors call pathways. The authors found that 75% of patients started on metformin. However, only 29% of patients remained on metformin monotherapy. The prevalence of persistent monotherapy varied dramatically across their data sources, from 10% to 80%. The most striking result from their study is the combinatorial number of potential treatment pathways observed.

While Hripcsak et al. quantify the diversity of diabetic treatment pathways, they fail to capture the time component of treatment evolution. Among the few studies that have considered time to treatment intensification is Berkowitz et al. [2014]. They analyze the transition from monotherapy to dual therapy. They sought to identify initial drug choices that are associated with an increased hazard of treatment intensification. However, intensification may occur many times during a patient's history. Berkowitz et al. [2014] only illustrate the first step of this process.

The challenge of identifying how patients traverse treatment choices shares complexity and motivation with clinical pathway discovery, as in Huang et al. [2013b] and Huang et al. [2013a]. However, while clinical pathway discovery seeks to learn movement through the clinical system, we are more focused evaluating how diabetes patients receive treatment given the existence of a solid, consistent system of guidelines. Recent work has offered new insight into how answering this type of question may proceed in practice [Yoon et al., 2013].

In the face of the little knowledge existing about treatment trajectories in practice, there is an opportunity to learn about the full trajectory of treatment intensification. This study defines a framework to begin this process. We focus on oral anti-hyperglycemic medication for treatment intensification and treat insulin initiation as a separate class of clinical outcome. The core structure for our work will be the count of oral anti-hyperglycemic medication that a patient takes over time. Particular medication choice often reflects individualized clinical decision making. To accommodate this, the guidelines couch their recommendations in overarching categories of monotherapy, dual therapy, and triple therapy, while making softer recommendations for drug choices within each category. We identify this flexibility as an asset. To allow for patient-specific treatment choices, we echo the guidelines and focus on drug count. In this framework, intensification is the addition of a medication, de-intensification the subtraction of a medication. We want to understand how intensification changes as a function of HbA1c measurements. The interplay between HbA1c and treatment intensification is relevant and nontrivial, with HbA1c being both the dominant trigger for intensification and the indicator for treatment effectiveness. Finally, we will consider insulin use as an escape from oral anti-hyperglycemic medication. That is, as soon as a patient begins using insulin, we will consider insulin use to proceed indefinitely. We recognize that different forms of insulin represent different treatment strategies, but we will treat all insulin initiation as exchangeable.

We use observational data to approach the problem of treatment trajectories from three directions. First, we report basic descriptive statistics of treatment patterns through medical claims and electronic health record data tracking 1 million diabetic patients from four national databases. Second, we look at the proportions of patients on different counts of oral anti-hyperglycemic over time after treatment initiation. Third, we model the number of oral anti-hyperglycemic medication a patient consumes as a Markov process, coloring drug count states with high (H) or low (L) HbA1c status, and using first insulin use as an absorbing state. We then fit this model to evaluate and compare intensification and de-intensification rates as functions of HbA1c level.

4.2 Methods

To learn about treatment in practice, we turn to four massive observational healthcare databases maintained by the Reagan-Udall Foundation Innovation in Medical Evidence Development and Surveillance (IMEDS) project. These databases contain time-stamped patient data, including laboratory results, drug exposures, and diagnoses, deidentified to compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA).

4.2.1 A diverse library of observational datasets

Three of these databases are from the Truven Health MarketScan Research family of datasets. One of these datasets, the MarketScan Commercial Claims and Encounters (CCAE) contains claims data from employees and their spouses and dependents covered by employer-sponsored private health insurance, including PPO and HMO plans, from inpatient and outpatient settings. The MarketScan Medicare Supplemental and Coordination of Benefits database (MDCR) includes inpatient and outpatient data from retirees with Medicare supplemental insurance paid by employers. This dataset includes Medicare-covered as well as employer or patient-covered expenses. The Medicare datasets have a history of use in understanding diabetes treatment use, as in Grabner et al. [2013], among others. Finally, the MarketScan Lab Database (MSLR) includes patients with inpatient and outpatient records, with an additional benefit of a high concentration of recorded laboratory results. We have access to one electronic health record (EHR) dataset as well. This is the General Electric Centricity Medical Quality Improvement Consortium (GECC) dataset. It contains ambulatory EHR data from providers using the GE Centricity record system who agree to share their data for research. Patients start their first medication from 2005 to 2014, from 2005 to 2014, from 2004 to 2014, and from 1994 to 2014 for the CCAE, MDCR, MSLR, and GECC datasets, respectively.

It is important to note the diversity of these datasets. We use the two major observational data modalities: claims data and EHR data. This is relevant because these data are not collected for the same purposes. The claims data reflect billing practices, while the EHR data do not. The EHR data are limited to clinical experiences where the EHR is deployed, while the claims data are limited by insurance acceptance policies. It is not immediately clear that one dataset is superior to another, but it is plausible that the sources of bias among them differ. This offers us the chance to broaden our perspectives of the problem at hand while lending credence to results shared among the datasets.

The development of a common data model (CDM) through the Observational Medical Outcomes Partnership (OMOP) experiment facilitates statistical methods implementation using these data. The CDM allows us to address pertinent questions about health practices, including comparative drug safety, by standardizing data concept representation across resources [Overhage et al., 2012]. Standardization takes the native representation in a clinical data set, such as International Classification of Diseases version 9 (ICD-9), of concepts like medication ingredients and diagnosis definitions, and translates them to a common representation. This facilitates consistent and reproducible analysis across datasets; we can apply the same analyses to each of these data resources, without having to recode our approach to accommodate dataset-specific variations. We depend on the OMOP CDM version 4 for representation of concepts of interest in all of our datasets.

One aspect of the OMOP CDM version 4 requires special attention. We use the OMOP CDM 4 definition of a drug era. A drug era is a combination of individual prescriptions or drug fills. For example, if the same medication is refilled routinely at the end of its 30 day supply for 2 refills, this appears as a single 90 day drug era. OMOP uses a standard 30 day persistence window, where if a new supply of the same medication is given within 30 days of the termination of a previous supply, it is considered the same era. For example, consider a patient who takes metformin for 60 days, forgets to refill a prescription for 4 days and does not take any medication. Then on the 5th day, that patient refills the prescription and continues taking metformin for 90 days. With a 30 day persistence window, all of the medication use actions result in a single 154 day drug era. The 30 day persistence window helps buffer refill discontinuities.

4.2.2 The study population

To create our study populations, we begin by selecting all patients who ever receive at least one diagnosis of T2D from each dataset, who are prescribed at least one oral anti-hyperglycemic medication, and who have at least two HbA1c measurements taken. Patients enter our study once they have had both a drug era started and a HbA1c measured. For CCAE, MDCR, and MSLR, 96% of patient visits were in the outpatient setting, with 3% in the emergency department (ED) and 1% in the inpatient setting. For GECC, 92% of visits were outpatient, 5% were inpatient, and 3% were in the ED.

To understand the clinical background of the patients, we collect age at first oral anti-hyperglycemic medication use, gender, HbA1c values, average serum creatinine and albumin:creatinine ratio (ACR) over all samples, and the number of statins ever used. We report the median, 25^{th}, and 75^{th} quartiles for each of these clinical parameters. We report the percent of patients who ever receive a diagnosis of cardiovascular disease (CVD), congestive heart failure (CHF), hypertension (HTN), hyperlipidemia, kidney disease, eye disease, neuropathy, or peripheral circulatory disorder. Although we rely on the OMOP CDM version 4 concepts, we select diagnosis concepts based on ICD-9 codes. For cardiovascular disease we used ICD-9 codes 410.*, 411.*, 412, 413.*, and 414.*. For congestive heart failure, we used 398.91, 402.*, 404.*, and 428.*. Hypertension and hyperlipidemia were identified with 401 and 272.*, respectively. Kidney disease corresponded to codes 581.81, 583.81, and 585.* . Eye disease corresponded to 362.07, 365.44, 366.41, and 369.*. Neuropathy consisted of 337.1, 353.5, 354.*, 355.*, 357.2,358.1, 396.54, 536.3, 713.5, and 782.0. Peripheral circulatory disorder related to 250.7, 443.81, and 785.4.

For our study, we will treat HbA1c as ≥ 7 (high) or < 7 (low), rather than as a continuous measurement. We do this to simplify the analysis. HbA1c of 7% reflects the value used in the UKPDS study and the ADA guidelines.

To have an overview of treatment trajectories, we extract summary statistics characterizing over-arching treatment patterns in each of our datasets. First, we consider the total time our patients are observable. We want to know how long patients persist in our dataset. We report the median observation time, where we define the median observation time as the total time each patient spends covered in our data. As a proxy for how frequently clinical interactions occur, we report median follow-up after the first treatment. We define follow-up time as the time between initiation of therapy and a relevant clinical event: a medication count change, a HbA1c measurement, insulin initiation, or any clinical visit. We count the proportion of patients who ever attain drug counts of 1, 2, 3, and 4 or more drugs during their time present in the dataset. We extend the persistence window concept to drug counts, where any pattern of de-intensification followed by intensification within 30 days is ignored, and we consider the count persistent. This buffers our data against mistaking medication switching with a de-intensification and intensification pattern. Finally, we report the proportion of patients who fail to change HbA1c status. These are the patients for whom all of their HbA1c measurements consistently fall within the high or low categories.

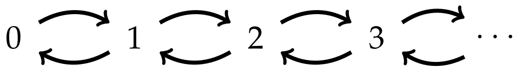
Although we will not consider medication type in our study, we query the percent of patients who have metformin as one of the treatments they receive within the first day of treatment. This allows us to compare our treatment initiation profile with that of Hripcsak et al.. Also, this helps orient us to how closely the patients we observe follow the strong recommendation to begin treatment with metformin.

We want to understand how patients progress through drug counts. To do this, we look at the percent of patients on 0,1,2,3, or 4 or more drugs at 3 month intervals following the initial treatment era beginning for 10 years. Not all patients persist for 10 years, and we show the percent extant patients from the starting cohort.

4.2.3 Birth-death processes

Birth-death processes have a significant history of use modeling populations [Kendall, 1948, Jaquette, 1970, Irvine et al., 1994]. When Kendall presents generalized solutions to the birth-death process, the problem is framed in the context of birth and death of individuals in a population [Kendall, 1948]. The population of interest varies with application. Early populations of interest included infected individuals in epidemics [Becker, 1972]. More recently, the populations of interest rely on clinical data. For example, Irvine et al. [1994] uses a modified birth-death process to model geriatric patients navigating in-patient resources. However, to the best of our knowledge, the application of birth-death processes to model patient treatment trajectories is novel.

A birth-death process is a Markov process operating on a discrete space over continuous time. State transitions only occur between neighboring states. Considering a simple birth-death process over integer values j, for a state j, one can only move from j to $j+1$ (a birth) or $j-1$ (a death). In many modeling situations, j represents the count of an item of interest. The facility of moving between states is captured by the transition rates, which measure the probability of moving from one state to another in infinitesimal time.



This simple model will form the skeleton of our approach. The states of interest are the counts of oral anti-hyperglycemic medications consumed. We extend this framework to capture some of the critical pieces to the clinical puzzle. First, we include the interplay between the treatment intensification and HbA1c. The guidelines make it clear that HbA1c should motivate clinical decision making; high HbA1c should drive drug counts higher or inspire faster progression to insulin use. However, increasing the number of diabetes treatments should reduce the disease impact, driving down HbA1c. It is possible to treat this interplay conditionally, looking at the count of drugs for patients with high or low HbA1c. However, this approach misses the feedback between HbA1c and drug treatments. We prefer to integrate HbA1c into the birth-death model. We allow each drug count state to assume both a high and low HbA1c level. This enables us to compare intensification / de-intensification rates between drug counts across HbA1c status.

In this model, we are assuming that HbA1c is fully known over time and stationary between observations, via the last observation carried forward. In reality, we only have partial observations of the complete process. However, we feel comfortable making this assumption because it represents the data

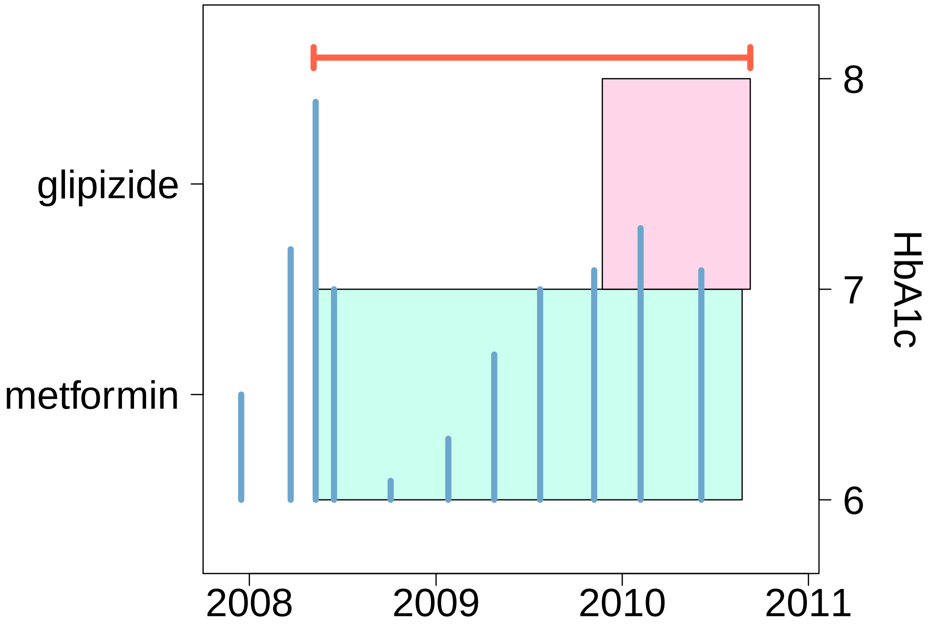


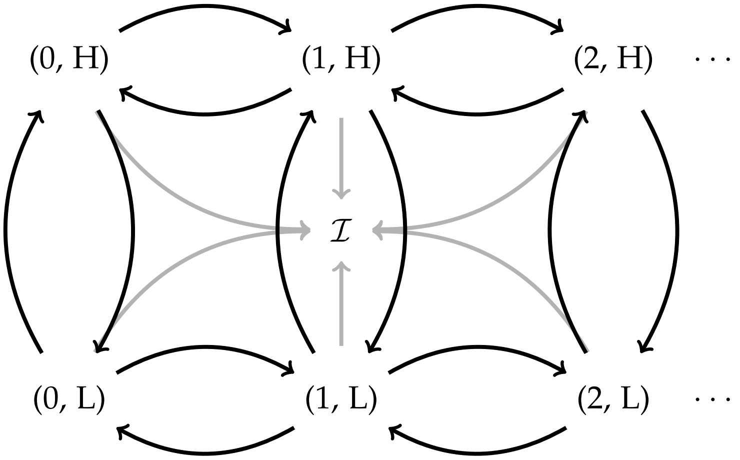
Figure 4.1: *The treatment trajectory for a 61 year old male patient from the MarketScan Lab database (MSLR) over the course of observation. Each rectangle represents a single drug era, with hemoglobin A1c (HbA1c) measurements shown as vertical bars. We mark the part of his history that enters into our study with a red segment above the measurement values.*

available to the physician. That is, in between observations it is likely that the physician refers to the last observed HbA1c. Therefore, we model the HbA1c between observations as equal to the last observed HbA1c.

Figure 4.1 shows a 61 year old male patient from the MSLR dataset. While the data tracks him from the first HbA1c measurement, he only enters into our birth-death model once both his HbA1c status is established and his first medication era begins. For him, this is shown with the red segment above his history. This structure forces us to ignore the time prior to the first known medication.

Second, we consider progression to insulin treatment separately from oral anti-hyperglycemic medication. Progression to insulin treatment is often non-reversible, with patients who begin insulin treatment often requiring insulin for the rest of their disease duration. With this in mind, we will model insulin initiation as an absorbing state. All treatment states can jump to the insulin use state, but no transitions exists from insulin.

Combining the insulin and HbA1c models together gives us our full model.



To keep track of this network of states, we need sufficiently general notation. Since we are not just moving along a linear chain of possible states, the notions of birth and death are less obvious. Let $\kappa_{m,n}$ be the transition rate from state m to state n. Let $\mathcal{D}_m$ be the set of possible destinations from m. If $m = (1, L)$, $\mathcal{D}_m = \{(0, L), (1, H), (2, L), \mathcal{I}\}$ and $\kappa_{(1,L),(2,L)}$ is the transition rate for escalating from one oral anti-hyperglycemic medications to two while remaining at low HbA1c. We define $T_{m,n}$ as the number of transitions from state m to state n and S_m as the total time spent in state m. Because we are considering patients as exchangeable and because the model is Markovian, we can sum all over all patients and observations for a single T for each pair of states and a single S per state.

4.2.4 Maximum likelihood estimates under the birth-death model

The likelihood under this model is well-known. Following Keiding [1975], among many others, the likelihood is

$$\mathcal{L} = \prod_{m}^{M} e^{-(\sum_{n\in\mathcal{D}_m} \kappa_{m,n})S_m} \prod_{n\in\mathcal{D}_m} {\kappa_{m,n}}^{T_{m,n}},$$

and the log likelihood is

$$L = \sum_{m}^{M} -(\sum_{n \in \mathcal{D}_m} \kappa_{m,n}) S_m + \sum_{n \in \mathcal{D}_m} T_{m,n} \log \kappa_{m,n}.$$

We can compute the maximum likelihood estimate for the parameters. Taking the derivative with respect to $\kappa_{p,q}$ we find

$$\frac{\partial L}{\partial \kappa_{p,q}} = -S_p + \frac{T_{p,q}}{\kappa_{p,q}}$$

Setting $\frac{\partial L}{\partial \kappa_{p,q}} = 0$ and solving yields intensification / de-intensification rate estimates

$$\hat{\kappa}_{p,q} = \frac{T_{p,q}}{S_p}.$$

To derive the asymptotic standard error, we compute the observed information matrix. The diagonal elements are readily found as

$$\begin{aligned} -\frac{\partial^2 L}{\partial \kappa_{p,q}^2} &= -\frac{\partial}{\partial \kappa_{p,q}} \left(-S_p + \frac{T_{p,q}}{\kappa_{p,q}} \right) \\ &= \frac{T_{p,q}}{\kappa_{p,q}^2}, \end{aligned}$$

while the off-diagonal elements simplify to

$$\begin{aligned} -\frac{\partial^2 L}{\partial \kappa_{r,q} \partial \kappa_{p,q}} &= -\frac{\partial}{\partial \kappa_{r,q}} \left(-S_p + \frac{T_{p,q}}{\kappa_{p,q}} \right) \\ &= 0. \end{aligned}$$

Inverting the observed information matrix and taking the square root gives the standard error. Thus, for $\hat{\kappa}_{p,q}$, the standard error is $SE(\hat{\kappa}_{p,q}) = \frac{\kappa_{p,q}}{\sqrt{T_{p,q}}}$. Armed with $\hat{\kappa}_{p,q}$ and its standard error, we can construct an asymptotic 95% confidence interval as $\hat{\kappa}_{p,q} \pm 1.96 \times SE(\hat{\kappa}_{p,q})$.

Table 4.1: *The demographic profile, including gender, age at the start of the study, comorbidities, important clinical outcomes, and other medication use of the diabetic patients included in the study from each of the datasets MarketScan Commercial Claims and Encounters (CCAE), General Electric Centricity Medical Quality Improvement Consortium (GECC), MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR), and MarketScan Lab Database (MSLR)*

	CCAE	GECC	MDCR	MSLR
n	114,060	757,135	27,073	133,042
Women (%)	46.6	51.7	48.3	47.1
Age (years)	55 [48,60]	61 [51,69]	71 [67,77]	56 [49,62]
Comorbidities				
CVD (%)	21.4	19.1	38.9	24.4
CHF (%)	6.7	7.5	20.2	9.3
HTN (%)	71.3	55.3	87.1	73.6
Hyperlipidemia (%)	88.1	79.2	91.5	88.5
Kidney disease (%)	9.5	13.0	35.2	14.1
Eye disease (%)	2.2	1.1	5.2	2.8
Neuropathy (%)	26.1	14.7	34.8	27.2
Peripheral circ. disorder (%)	1.0	0.5	5.9	1.9
Clinical outcomes				
Hypoglycemia (%)	3.2	1.7	4.6	3.5
HbA1c	6.9 [6.3,7.9]	6.8 [6.2,7.6]	6.8 [6.3,7.5]	6.9 [6.3,7.8]
Serum creatinine	0.9 [0.7,1]	1 [0.8,1.2]	1 [0.8,1.3]	0.9 [0.8,1.1]
ACR	7 [4,18]	NA	10 [5,30]	7 [4,20]
Other medications				
Statins used	1 [1,2]	1 [1,2]	1 [1,2]	1 [1,2]

4.3 Results

We extract study populations from each of the four national databases, yielding in total over 1 million T2D patients with at least one oral anti-hyperglycemic claim or prescription and at least two HbA1c laboratory measurements. Table 4.1 offers a summary of the patient demographics and their health status.

4.3.1 Study population demographics

We begin with the general trends of drug use among patients, showing our results in Table 4.2. Patients are present in our datasets for a median of 5 to 6 years. This reflects the entire time they are eligible for recording events.

Table 4.2: *A clinical treatment profile of the diabetic patients included in the study from each of the datasets CCAE, GECC, MDCR, and MSLR. We show the median observation length in the dataset and follow-up time from the onset of treatment, where follow-up was defined as a clinical visit, a diabetes medication count change, a HbA1c test, or progression to insulin. We report the percent of patients who start treatment with metformin. We compute the percent of patients who ever reach a given count of drugs, independent of time, in our data.*

	CCAE	GECC	MDCR	MSLR
Median observation time (y)	5.9	5.1	5.1	5.6
Median follow-up time (d)				
After first drug	20	30	16	19
Starting treatment (%)				
Metformin	76.8	75.4	63.5	74.8
Proportion ever attaining drug count (%)				
1 drug (%)	83.7	91.9	86.4	84.5
2 drugs (%)	53.2	52.4	48.8	52.8
3 drugs (%)	20.1	19.4	14.9	19.5
4+ drugs (%)	3.9	3.8	2.2	3.7

For all datasets, the overall median follow-up time from initiation of the first treatment varied from 15 to 30 days. We defined the follow-up time as either a relevant diabetes clinical event, including HbA1c measurement and medication change, or a clinical visit for any reason. Our rational for these choices is that each of these events represents an opportunity for physician-patient interaction regarding treatment.

We also report an overview of treatment decisions. We find that roughly 75% of patients begin treatment with metformin, with the exception of the MDCR data, where only 63.5% of patients begin treatment with metformin. Additionally we examine the proportion of patients ever attaining a number of medications. Specifically we count all patients who are on a medication for at least one day at any point during their observation period. Roughly 50% of patients will take two drugs concurrently at some point, and roughly 20% ever take three drugs concurrently. However, the proportion of patients on 4 drugs or more is less than 5%.

We also examine the dynamics of HbA1c testing across the datasets. In Table 4.3, we see the median number of observations per person, the median period of HbA1c testing, and the median time in between tests. Each patient

Table 4.3: *The hemoglobin A1c (HbA1c) testing profile of the diabetic patients included in the study from each of the CCAE, GECC, MDCR, and MSLR datasets. We report the time in days. For each quantity, we show the median and upper and lower quartiles.*

	CCAE	GECC	MDCR	MSLR
Count HbA1c / person [median (Q_1,Q_3)] (%)	4 [2,6]	6 [3,11]	5 [3, 7]	4 [2,6]
HbA1c testing period median (Q_1,Q_3)] (%)	472 [273, 803]	223 [155,354]	365 [212, 638]	410 [247, 711]
Time between tests [median (Q_1,Q_3)] (%)	134 [92, 220]	129 [95,195]	139 [92, 209]	135 [92, 218]

has, on average, 4 to 6 tests over her time in the study, and we have already selected this population so that all patients have had at least two HbA1c tests. The median period examines how many days occur in our study for each HbA1c test. For all of the datasets, this ranges between half a year to a year and a half per test. Similarly, we report the median time between tests. This is the average time from one test to another, which differs from the median period by not considering time between the study start and the first HbA1c test or the time between the last HbA1c test and the study end.

We next turn to with the general trends of HbA1c dynamics among patients, showing our results in Table 4.4. A remarkable result is the proportion of patients who remain at either a high or low HbA1c state throughout the duration of their observation. Collectively, across all datasets, between 49% and 64% of patients remain in their initial HbA1c state, independent of treatment. Between 20% and 30% of patients remain in a high HbA1c state independent of treatment. Similarly, between 29% and 40% of patients remain at a low HbA1c state. Within the high HbA1c group, the median A1c values are 8.4, 8.1, 7.9, and 8.3, and in the low HbA1c group, the median A1c values are 6.1, 6.1, 6.2, and 6.1 for CCAE, GECC, MDCR, and MSLR, respectively.

We next turn to the relative breakdown of HbA1c events among patients who start with high or low HbA1c. Specifically, we report the percent of patients who start at high or low HbA1c and then the percent who move from high to low, or high to low and back again, aggregating the group that makes at least three such transitions. We show these results in Table 4.4. Between 40% and 60% of patients who start with high HbA1c never achieve low HbA1c,

Table 4.4: *These are a description of the HbA1c events among patients who start with high or low HbA1c. We show the relative number of patients who remain stationary at their HbA1c category, as well as the most frequent movement patterns. The percent of patients starting at high versus low HbA1c is remarkably similar. Relatedly, all of the datasets have largely consistent proportions of each transition type.*

	CCAE	GECC	MDCR	MSLR
n	114,060	757,135	27,073	133,042
Start ≥ 7 (%)	49.1	50.2	58.8	50.7
Always ≥ 7 (%)	29.5	20.2	21.5	28.0
$\geq 7 \rightarrow < 7$ (%)	8.2	8.1	9.2	8.4
$\geq 7 \rightarrow < 7 \rightarrow \geq 7$ (%)	3.7	5.3	5.6	4.1
Other	2.9	7.7	4.5	3.3
Start < 7 (%)	50.9	49.8	41.2	49.3
Always < 7 (%)	34.3	29.1	39.5	35.0
$< 7 \rightarrow \geq 7$ (%)	12.5	12.5	10.6	12.2
$< 7 \rightarrow \geq 7 \rightarrow < 7$ (%)	5.5	7.3	5.1	5.5
Other	3.4	9.8	4.0	3.6

even transiently. In contrast, less than 20% make exactly one transition from high HbA1c to low HbA1c. Relatively far more patients make the transition from low HbA1c to high HbA1c one time.

4.3.2 Trends in medication count over time

Assessing the percent of people at given drug counts at three month follow up intervals post initial treatment reveals that the distribution of medication counts among the study patients changes rapidly in the first year, supplanted by much more gradual change thereafter. We report these distributions in Figure 4.2. All datasets show a similar rapid descent of the population of patients taking only one drug, with an attendant rise in the percent of patients taking zero, two, or three drugs. The percent of patients remaining on one drug hovers around 40% for all datasets. For all but the GECC dataset, the next highest population is the group of patients taking two drugs, whose proportion increases from roughly 20% within the first year toward 30% to 40% at 10 years. In the GECC dataset, the patients taking zero drugs dominate, but their population steadily declines from a peak near 40% within the first year after initial treatment initiation. The absolute number of patients observed at each sample is strictly decreasing with a super linear change.

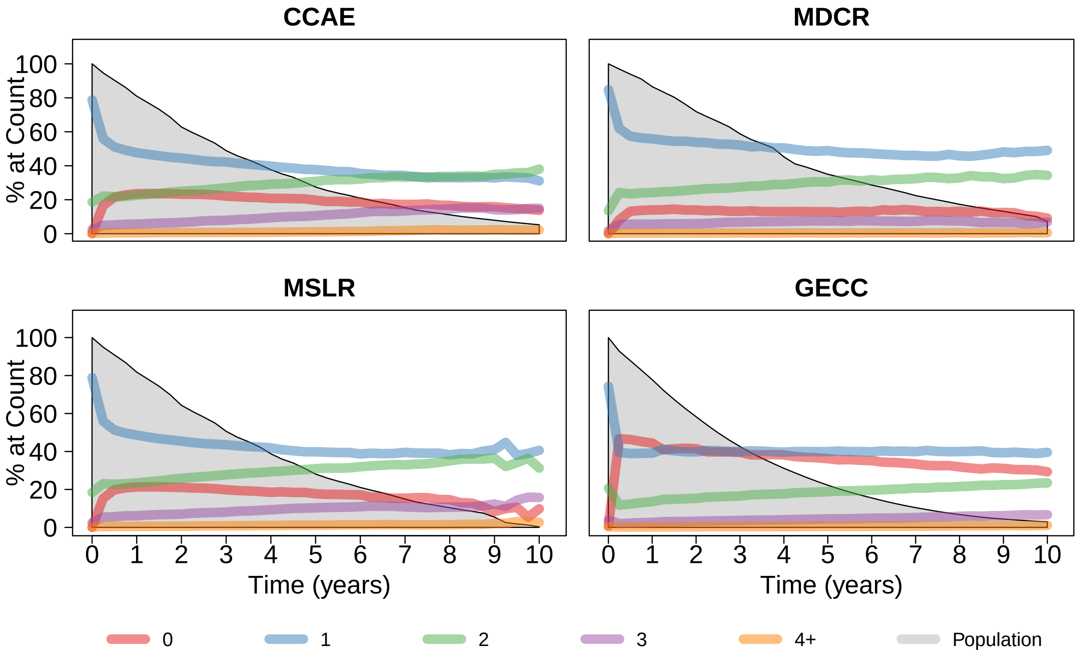


Figure 4.2: *Tracking the distribution of patients at given counts of drugs over time after the beginning of their first eras of treatment suggests a similar trajectory across each of the MarketScan Commercial Claims and Encounters (CCAE), MarketScan Medicare Supplemental and Coordination of Benefits database (MDCR), MarketScan Lab Database (MSLR), and General Electric Centricity Medical Quality Improvement Consortium (GECC) datasets. In all datasets, the percent of patients on one drug falls dramatically within one year, replaced by patients taking 0 or 2 drugs. After the initial recalibration, the proportion of patients at each count changes slowly for the rest of observation.*

4.3.3 Full treatment trajectories

Turning to the results from our birth-death model, Figure 4.3 provides the first full depiction of the treatment trajectory including HbA1c and insulin initiation. The time spent in each state is proportional to the node size in our graph. For the intensification / de-intensification rates, the edge thickness is proportional to our maximum likelihood rates. Edge opacity is inversely proportional to the standard error of the rates; rate estimates of which we are more certain appear darker.

From the figure, we see that the time spent on 0, 1, and 2 drugs dominates all the datasets, with time spent on 3 and 4 or more drugs appearing relatively small. For most of the count states, the time spent in the low HbA1c category appears higher than the time spent in the corresponding high HbA1c category. Except for the zero to one transition, de-intensification rates appear higher than intensification rates, but the intensification and de-intensification rates dwarf the transition rates between the high and low HbA1c states and the transition rate to insulin initiation.

Focusing on the patients who fail to change HbA1c status shown in Table 4.4 we recreate the plots from Figure 4.3, restricting ourselves to the patients with perennially high or low HbA1c, which we treat separately. The resulting plots are shown in Figure 4.4. Using the birth-death modeling framework, we compare the 95% confidence intervals we constructed between the always high and always low corresponding transition rate pairs for each dataset. A red asterisk labels each of the transition arrows that is significantly higher than its counterpart in the other group from the same dataset. Except for three transitions in the GECC dataset, all of the asterisks label the high HbA1c group of patients for each of the datasets. That is, for most intensifications, de-intensifications, and transitions to insulin, the transition rates computed using the population of patients whose HbA1c values are not seen falling below 7% are significantly higher than the corresponding rates for the patients whose HbA1c values are not seen rising above 7%.

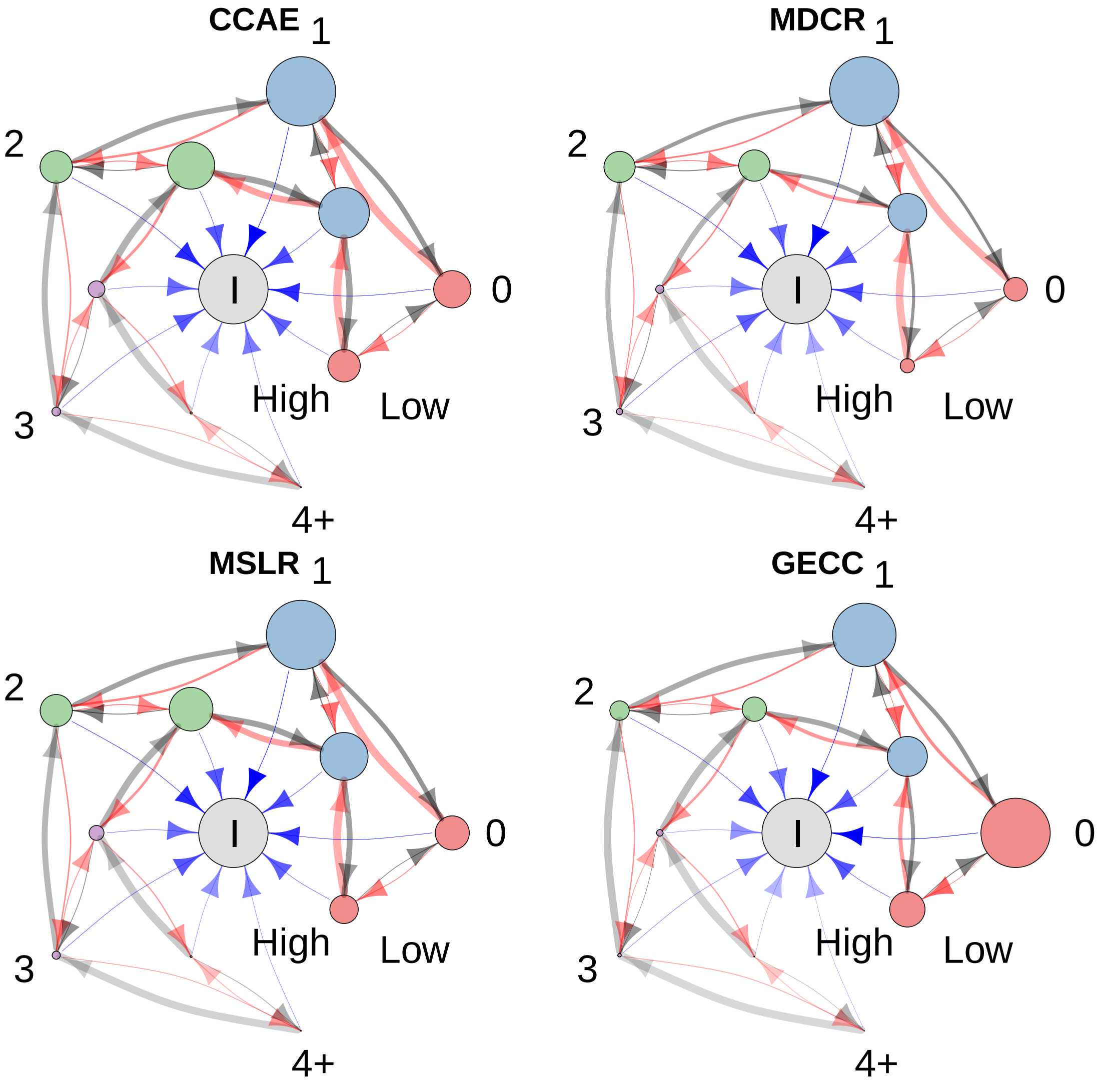


Figure 4.3: *Birth-death processes including high and low HbA1c status and insulin initiation (I) for patients in the CCAE, GECC, MDCR, and MSLR datasets. Each of the drug count states is a circle connected by arrows that reflect the rate of transition between states. The size of the nodes is proportional to the time spent in drug count state,* $\{0, 1, \ldots, 4+\}$*, the edge widths are proportional to the maximum likelihood transition rates, and opacity is inversely proportional to standard error. Drug counts with high HbA1c status lie in the inner circle; those with low HbA1c status remain further away from the central insulin node.*

The time spent on 0, 1, and 2 drugs dominates all the datasets, with time spent on 3 and 4 or more drugs appearing relatively small. For most of the count states, the time spent in the low HbA1c category appears higher than the time spent in the corresponding high HbA1c category. De-intensification rates often appear higher than intensification rates. Most notably, the intensification and de-intensification rates dwarf the transition rates between the high and low HbA1c states and the transition rate to insulin initiation.

4.4 Discussion

This research provides an emerging depiction of patient trajectories through T2D treatment. Given the challenge of observational data, recovering patterns that we expect can validate our conclusions. Beginning by examining our patient cohort, we see trends that we naively expect. Considering Table 4.1, the MDCR patients are generally older and have more comorbidities than their counterparts from the other datasets. We expect medicare patients to represent an older population compared to patients who have insurance through employers. Furthermore, it is reasonable that older patients have more comorbidities on average than their younger counterparts.

The percent metformin use at treatment initiation is largely consistent with previous findings [Hripcsak et al., Berkowitz et al., 2014, Grimes et al., 2015, Weng et al., 2016]. The question remains why many patients begin on a drug other than metformin. One reason might be that metformin is contraindicated for some patients. Another reason might be treatment initiation outside of our datasets. The short median observations times suggest that much of the patient history exists outside of these datasets. Ultimately, this is a shortcoming of decentralized claims and EHR data, where patients may pass in and out of the isolated systems.

While our datasets contain inpatient, outpatient, and ED visit events, the overwhelming majority of visits occur in the outpatient setting. This is reassuring, since treatment decisions in the inpatient or ED settings may differ from standard, long-term care approaches as an outpatient. Also, this underscores the relevance of our results to clinicians working in an outpatient setting.

One of our most surprising findings is that there is a low level of transition between the high and low HbA1c states. Half to two-thirds of the patient population fails to change HbA1c status. This is unexpected for several reasons. First, the population of patients at high HbA1c for the duration of their observation suggest that treatment with oral anti-hyperglycemic medication may be ineffective at dropping the HbA1c of some patients below the 7% threshold. Conversely, the patients who remain below 7% suggest that a population exists whose disease course may be less severe, or for whom treatment may inhibit disease course. Our estimates for low transition rates from high to low HbA1c are not isolated [Maclean et al., 2004].

Comparing the high and low HbA1c populations, intuition suggests that the transition rates for intensification should be greater in the high HbA1c population, and we recover this result. None of the intensification rates in the low HbA1c population significantly exceed the corresponding rates in the high HbA1c population. Similarly, the rates of transition to insulin in the low HbA1c population never significantly exceed their counterpart rates in the high HbA1c population. However, the de-intensification rates are also often higher for the high HbA1c population.

High intensification and de-intensification rates together suggest that the treatment states may be less stable for the high HbA1c patients. We could be seeing the results of patients trying different medications. By moving onto and off of different medication, patients would appear to have transient intensification and de-intensification events. It is possible that we are seeing issues of adherence. Patients who are poorly compliant would have a propensity to both stop and restart medication regimens and have higher HbA1c. These explanations are not mutually exclusive. In fact, our results may reflect a cycle of frustration, absence of clinical response, and poor adherence, as frustrated patients with high HbA1c that is resistant to treatment become less compliant.

Measuring adherence is challenging, but it has a notable impact on HbA1c [Lawrence et al., 1970, Pladevall et al., 2004]. We ultimately do not know with certainty when patients took medication, even if it was prescribed. This is an inherent limitation of working with claims and EHR data. Future work should create an independent compliance proxy and evaluate how that compliance changes as a function of time for high HbA1c patients.

Poor follow up and compliance may not present the whole story. In particular, the median follow up times across the datasets are quite low. This suggests that physicians and patients are communicating. While some of these events are certainly not diabetes related, the frequency of follow up suggests that physicians have the opportunity for intervention on patient treatment trajectories. Phillips et al. [2001] discusses the phenomenon of clinical inertia, where clinical treatment aggressiveness may be dampened in conditions where treatment is not symptom motivated. Citing Harris et al. [1999], Phillips et al.

[2001] report that only 33% of diabetic patients are adequately treated to lower HbA1c below 7%, echoing our observed percent of patients who remain at high HbA1c. This phenomenon has appeared before [Dailey et al., 2002, Slabaugh et al., 2015].

Observational data are notorious for producing different results for different databases [Madigan et al., 2014, 2013]. When different databases, with different populations, show similar results, it is worth noting. Here, we see similar trends for all the databases in Figure 4.2. After treatment initiation, all datasets show that the percent of patients on a given medication count are similar over time. Furthermore, the plateau level of patients on monotherapy hovers between 40% and 60%, echoing the results of Alexander et al. [2008], Weng et al. [2016]. Similarly, our reported levels of duel therapy may seem low. However, the levels we see are very close to other studies; Qui et al. [2015, 2012] report similar trends, despite looking exclusively at new users.

Clinically, this is really a cause for concern. The UKPDS study demonstrates that 50% and 75% of patients will need an additional treatment to manage their diabetes at 3 and 9 years, respectively [Turner et al., 1999]. The percent of patients actually taking at least one additional medication at 3 and 9 years after treatment initiation falls woefully short of 50% and 75%. This suggests that clinicians may need to be more willing to intensify treatment than they show in practice.

Insulin use rates are low. Compared to the rates of oral anti-hyperglycemic medication, the rates of transition to insulin are markedly less. One obvious explanation for this reticence is that there is a preference for avoiding insulin on the part of the physicians and patients. 10-15 years used to be the standard wait time between onset to insulin use, [Nathan, 2002]. Current guidelines suggest at most 9 months between treatment initiation and insulin [American Diabetes Association, 2015].

Working with observational data is inherently problematic, and many of limitations apply here. First, all records relevant to each patient may not be present in the dataset. For example, the GECC data reflects the prescriptions written by a majority population of primary care physicians [Crawford et al., 2010]. If a patient were to transfer care to a specialist, that patient might seem to disappear from our records. If this patient then reenters our dataset after some time under the care of a specialist, it would seem that there were on no treatments for a long time, when the reality may be far different. The shadow of this limitation is visible in Figure 4.2. Specifically, in Figure 4.2, a much larger percent of patients persist at 0 drugs over time in the GECC population. This is possibly due to patients leaving the system for specialist care. However, the percentages of patients remaining on monotherapy versus discontinuing therapy actually have precedent in the literature [Hazel-Fernandez et al., 2015]. Some have offered economic reasons for departure from guidelines [Vigersky et al., 2013].

Similarly, frequency of HbA1c testing depends on physician and patient, and inadequate testing is not uncommon [Maclean et al., 2004]. Decisions to test more or less frequently might alter our ability to draw conclusions from this data. Future work should look closely at the relationship between going on and off of a drug and the frequency of HbA1c testing and should be willing to remove patients from our study who are insufficiently tested.

While it is concerning that the majority of patients appear to stop treatment at some time, treatment changes are common [Boccuzzi et al., 2001]. This is possibly a reflection of prescription era modeling decisions. The 30-day persistence window is an arbitrary interval. It is still possible to observe disjoint eras, when a single continuous era would be more clinically reasonable. Medication may be taken sporadically, with some doses stored and then used later. Future work should investigate how clinical information can better be integrated into these drug eras.

Modeling the count of drugs introduces notable issues during medication switching. In particular, if a patient changes from one medication to another, this can enter the data as an erroneous intensification or de-intensification. For example, if the patient switches midway through a prescription and receives a secondary prescription, the overlapping eras will appear as a transient escalation to two drugs. Unfortunately, these issues are difficult to avoid in an observational setting, when the only data we see are drug eras.

Another limitation of this study is the transformation of the HbA1c data from a partially observed continuous process into a completely observed discrete process. The true HbA1c for a given individual moves stochastically between observations, and it would be more meaningful to learn about this process including the drug counts during the unobserved period. However, we justify our approach by recognizing that we are in some sense modeling the physician's decision making, and it would be reasonable for a physician to make treatment intensification decisions reliant on the last observed HbA1c rather than the estimate of the current HbA1c. Future work will include identifying the effect and importance of HbA1c cutoff and granularity. It is reasonable to suspect that physician decision making as a function of HbA1c is far more nuanced than branching on a high or low value.

Although the limitations of observational data are not to be underestimated, one of the goals of this research is to highlight practice in a realistic clinical setting. The questions of compliance and data completeness in our data resemble the uncertainty of clinical decision making. Thus, in some sense, the limitations of observational data force us to account for some of the same uncertainties facing clinical decision makers. Our approach is a first step to understanding the treatment trajectories of patients, offering insight into how patients are treated and framing the conversation in terms of trajectories for the first time.

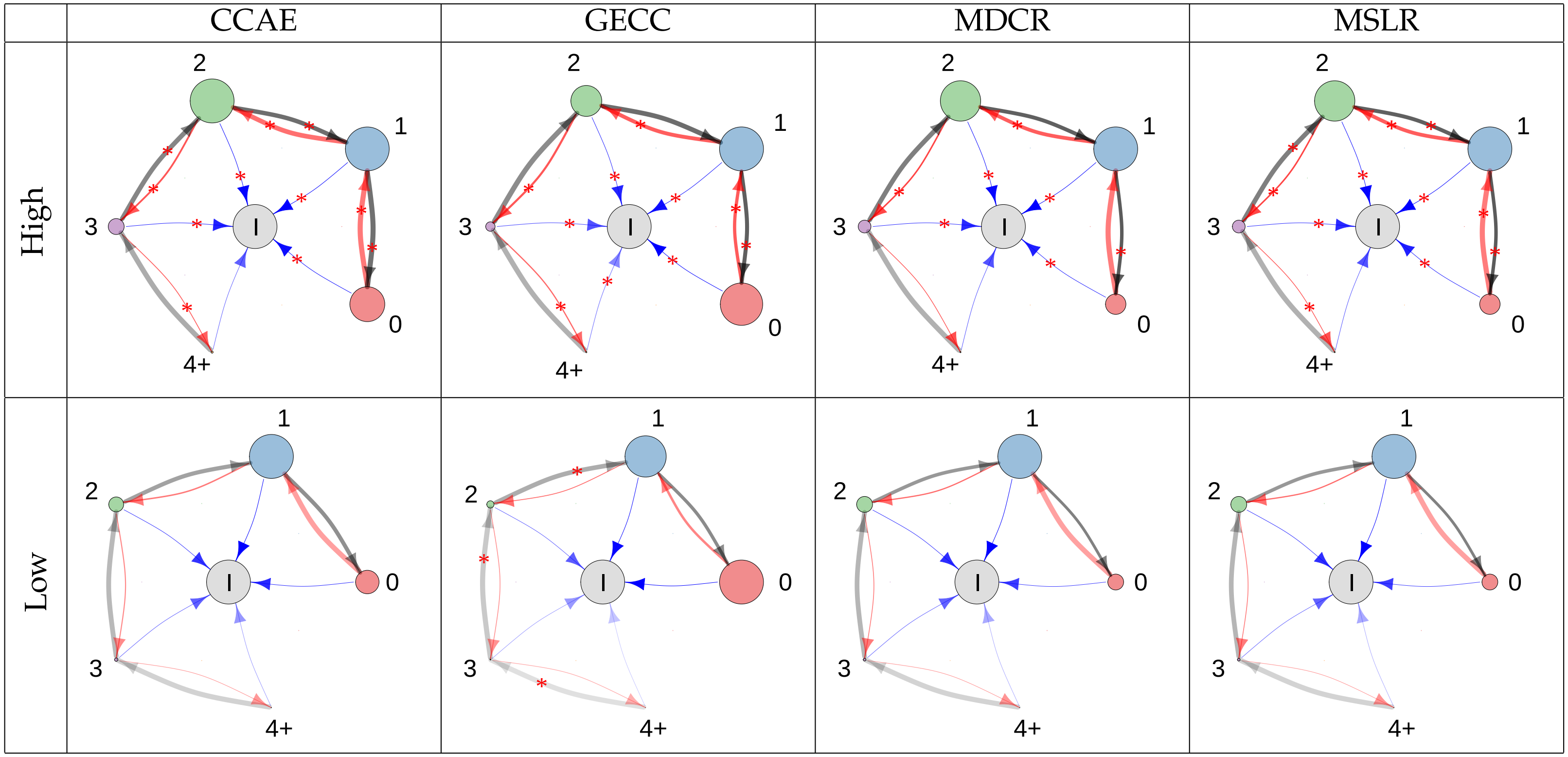


Figure 4.4: *Comparing the trajectories of patients who remain within the high or low HbA1C in the study populations. The high HbA1c population estimates appear in the top row, and the low HbA1c population estimates appear in the bottom row. Populations from the same dataset share a column. A red asterisk indicates that the transition rate is significantly larger than the counterpart transition rate from the other population.*

CHAPTER 5

Future Work

5.1 Markov chain Monte Carlo for SCCS at scale

Bayesian ideas have won approval in the setting of observational healthcare data. Placing prior distributions over the estimates of relative risk for each drug captures our belief that most drugs are safe and allows us to consider related outcomes. In addition, using a Bayesian framework efficiently manages challenges in estimation at scale [Madigan et al., 2011]. However, moving from posterior mode estimates to extracting full posterior distributions is a standing challenge in this setting.

5.1.1 Motivation

Current approaches for capturing uncertainty of our risk estimates for medical interventions rely on bootstrapping. In bootstrapping, our goal is estimating the standard error of a parameter of interest Efron and Gong [1983]. Bootstrapping is a non-parametric approach. We approximate the standard error of a parameter by estimating that parameter repeatedly using data resampled from the full dataset with replacement. Although computationally intensive, it is not prohibitive in the setting of observational healthcare data. In particular, Suchard et al. [2013] bootstrap the mode estimates.

However, this approach is notoriously problematic. The Bayesian modeling that we use is equivalent to a regularized regression framework, where our prior distribution results in the penalization term. One particular example of this is a Lasso penalized regression [Tibshirani., 1996]. The Lasso estimator optimizes an l_1 penalized regression. This regression corresponds to using a Laplace prior in our formulation. There are two main benefits to using a Lasso penalized regression. First, it enforces sparse solutions, shrinking small estimates to zero and producing a parsimonious model. Effectively, this is simultaneous model selection and model fitting. Second, it allows for computationally feasible inference [Tibshirani., 2004].

This shrinking effect is problematic when combined with bootstrapping. Chatterjee and Lahiri [2010] consider bootstrapping a Lasso penalized linear regression. They find that the bootstrap estimator converges weakly to a random probability measure, rather than the target distribution. In particular, their results show that the bootstrap is inconsistent when regression coefficients are forced to zero. This is precisely the setting in which we expect our Lasso penalized regressions to live. Chatterjee and Lahiri [2011] offer a modified bootstrap approach to correct for this phenomenon.

However, a similar problem arises from selecting the hyperprior. To select the hyperprior, the current best practice is to use cross-validation [Suchard et al., 2013]. We use cross-validation based on the predictive log-likelihood of the hold-out set to select the prior variance $\sigma^2 = \frac{1}{\tau}$. Suchard et al. [2013] use a log-scale grid search that is computationally expensive even with only a single parameter. To help overcome this burden, we turn to Genkin et al. [2007] in implementing an “autosearch" for hyperparameter selection. We start with an initial guess and then increase or decrease our guess by one log unit until we have bracketed the maximum of the hold-out set predicted log-likelihood. Then we compute a quadratic approximation to the predicted log-likelihood. The maximum of this approximate surface becomes our estimate.

Ultimately, both the bootstrapping approach for estimator variability and the predictive log-likelihood cross-validation approach for estimating the prior variance fall short of capturing what our Bayesian framework really demands: the full posterior distribution. In this project, we address these shortcomings by developing fully Bayesian inference for SCCS using our massive datasets. We address the challenges of dimensionality by exploiting model averaging. We first simulate from the marginal posterior distribution of covariate inclusion using a Laplace approximation. Next, we use Metropolis-within-Gibbs to learn about both the drug relative risks given each covariate inclusion model and the hyperprior variance. In the Metropolis step, we develop an adaptive independence sampler with proposals from a tuned multivariate normal distribution around the mode estimate.

5.1.2 Methods

5.1.2.1 SCCS

In using the SCCS model, we follow the notation of Simpson et al. [2013] and Suchard et al. [2013]. To revisit our nomenclature, the SCCS model assumes that ADEs arise according to an inhomogeneous Poisson process. For $j = 1 \dots J$ drugs under consideration, the parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)'$ measure the instantaneous log relative risks of treatment exposure. As before, let patients $i = 1, \dots, N$ have a baseline risk e^{ϕ_i} and let the drug exposures multiplicatively modulate the underlying instantaneous event intensity λ_{ik} during constant drug exposure era k. That is, the intensity arises as $\lambda_{ik} = e^{\phi_i + \mathbf{x}'_{ik}\boldsymbol{\beta}}$, where $\boldsymbol{x}_{ik} = (x_{ik}, \dots, x_{ik})'$ and x_{ikj} indicates exposure to drug j in era k for outcome p. The exposure duration for exposure era k of patient i is l_{ik}. The number of ADEs in era k of patient i is $y_{ik} \sim \text{Poisson}(l_{ik} \times \lambda_{ik})$. The SCCS method conditions on the total number of events for a particular outcome $n_i = \sum_k y_{ik}$ that a patient experiences over her total observation period. By conditioning on these statistics, the baseline risk falls out of the conditional likelihood of the data.

We place a prior distribution over each of the covariates

$$\begin{aligned} p(\boldsymbol{\beta}|\sigma^2) &\sim \prod_j \left[\text{Normal}\left(0, \frac{1}{\sigma^2}\right)\right] \\ p(\boldsymbol{\beta}|\sigma^2) &\sim \prod_j \left[\text{Laplace}\left(0, \frac{1}{\sigma^2}\right)\right]. \end{aligned} \tag{5.1}$$

As the model is currently formulated, we select the prior variance σ^2 that maximizes out of sample prediction through cross-validation. However, we would like to place a distribution over this hyperprior and learn its value. We start by making this adjustment to the hyperprior modeling. We place an inverse-gamma distribution over our hyperprior variance σ^2. This approach is commonly used [George and McCulloch, 1993].

5.1.2.2 Spike and slab prior

Variable selection is a long standing challenge [Ishwaran and Rao, 2005]. Some approaches offer theoretically perfect model selection by considering all 2^J models for J covariates, as compared in Shao [1997] among others. However, these methods may fail in practice [Shao and Rao, 2000]. Furthermore, enumerating 2^J for moderate J quickly become impractical.

One of the Bayesian solutions to this problem was the development of spike and slab prior distributions [Mitchell and Beauchamp, 1988, George and McCulloch, 1993]. Mitchell and Beauchamp [1988] designed the spike and slab prior as a tool for selecting a subset of variables within a model. The core idea of the spike and slab prior is a hierarchy of prior distributions over the parameters and the model [Ishwaran and Rao, 2005]. Some model parameters are vulnerable to exclusion, and that we would like the data to choose which of these variables to remove. As such they allow covariates to have discrete probability mass at zero. This discrete mass represents the “spike" component of their name. Functionally, this component of the prior hierarchy is responsible for model selection.

However, the non-zero covariates would still have a prior distribution over them. Mitchell and Beauchamp [1988] maintain a diffuse prior distribution over these non-zero values as well. This diffuse component of the prior is the “slab" part of the name. This component is already built into the SCCS model; the Normal and Laplace priors we currently use are functionally “slab" priors. Our current model formulation is a degenerate case of the spike-and-slab prior model - one without any spikes.

In the setting of observational healthcare data, using priors of this form is reasonable. First, our prior distributions over the covariates are already centered at zero, reflecting our belief that most diagnosis-intervention relationships are null. Therefore, placing a point mass over zero merely underscores this belief. Second, for any given adverse event, it is reasonable that most drugs will be unrelated to it.

Given some shape and scale parameters κ and θ for our hyperprior distribution, our prior structure is now

$$\sigma^2 \sim \text{Inv-Gamma}(\kappa, \theta) \tag{5.2}$$

and

$$p(\boldsymbol{\beta}|\sigma^2) \sim \prod_j \left[\delta_0(\beta_j) + \text{Normal}\left(0, \frac{1}{\sigma^2}\right)\right] \tag{5.3}$$

or

$$p(\boldsymbol{\beta}|\sigma^2) \sim \prod_j \left[\delta_0(\beta_j) + \text{Laplace}\left(0, \frac{1}{\sigma^2}\right)\right]. \tag{5.4}$$

Note our use of $\delta_0(\beta_j)$, a delta function over β_j to represent the point mass at 0 for β_j. To capture our spike and slab prior, we consider the set of models $\mathcal{G} = \{g_0, g_1, ...g_{2^J}\}$, where g_j represents the set of non-zero covariate values. There are 2^J possible models.

Gibbs sampling is commonly employed to learn about models with spike and slab priors [George and McCulloch, 1993]. We follow the spirit of this approach. But, to learn about our models, we use the Laplace approximation to the posterior distribution, developed by Tierney and Kadane [1986]. In the Laplace approximation, we use Laplace's method to approximate our posterior density. Laplace's method relies on a second order Taylor expansion about the posterior mode. Specifically, we find that the posterior is approximated by a normal density centered at the posterior mode with a covariance equal to minus the inverse hessian at the mode.

Armed with our cyclic coordinate descent framework, we have the tools necessary for the Laplace approximation already in hand. Specifically, we already compute the posterior mode $\hat{\boldsymbol{\beta}}$. Computing the Hessian at $\hat{\boldsymbol{\beta}}$ is straightforward. Under the SCCS model, the Hessian at the mode of model j is

$$\frac{\partial^2 l(\boldsymbol{\beta}_j)}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j} = \sum_i^N n_i \left[\left(\frac{\sum_{g=1}^{G_i} t_{ig} e^{\boldsymbol{x}'_{ig}\beta_j} \boldsymbol{x}_{ig}}{\sum_{g'=1}^{G_i} t_{ig'} e^{\boldsymbol{x}'_{ig'}\boldsymbol{\beta}_j} \boldsymbol{x}_{ig'}} \right)^{\otimes 2} - \frac{\sum_{g=1}^{G_i} t_{ig} e^{\boldsymbol{x}'_{ig}\beta_j} \left(\boldsymbol{x}_{ig}\right)^{\otimes 2}}{\sum_{g'=1}^{G_i} t_{ig'} e^{\boldsymbol{x}'_{ig'}\boldsymbol{\beta}_j}} \right], \tag{5.5}$$

where $\otimes$ is the Kronecker product. Our covariance matrix for the approximating normal distribution is $\hat{\mathbf{\Sigma}}_j$. We construct

$$\hat{\mathbf{\Sigma}}_j = -\left(\frac{\partial^2 l(\boldsymbol{\beta}_j)}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j}\right)^{-1}. \tag{5.6}$$

Therefore we approximate the posterior distribution with $N(\hat{\boldsymbol{\beta}}_j, \hat{\mathbf{\Sigma}}_j)$.

Exploring the space of $\mathcal{G} = \{g_0, g_1, ...g_{2^J}\}$ is straightforward. We move through the space of models by introducing or removing one variable at a time. If we consider the current model g_j and propose model g_k, we accept g_k with probability

$$p = \frac{L_{g_k}\pi_k}{L_{g_j}\pi_j} \tag{5.7}$$

where

$$L_{g_k} = \frac{L(\hat{\boldsymbol{\beta}}_k)}{(|\hat{\mathbf{\Sigma}}_k| 2\pi)^{\frac{1}{2}}} \tag{5.8}$$

is the Laplace approximation likelihood and π_j is the prior contribution for model g_j.

5.1.2.3 Metropolis-within-Gibbs

We will use Markov chain Monte Carlo (MCMC) to construct our posterior distribution. One of the great challenges of MCMC in high-dimensional problems is poor mixing, or the slow convergence of an MCMC chain to a stationary distribution. Many have tried to find solutions to this problem [Roberts and Rosenthal, 2009]. However, this remains a challenge. For us, this is critical when we are considering implementing MCMC to learn about models with thousands to tens of thousands of parameters. Also, it is frustrating to be limited by this dimensionality when we strongly believe that most parameters will have negligible relevance for a given outcome of interest. By selecting smaller models through the spike and slab prior, we mitigate some of the dimensionality issues.

We will average over the models selected from the Gibbs-like process above. Specifically, we will record the frequency with which each model is visited. Then we will learn about the parameters included in a model given that model.

Many Markov chain approaches are available to sample from our posterior distribution. Two of the canonically dominant methods are the Metropolis-Hastings algorithm and Gibbs sampling. While the method of choice for exploring the posterior space should not alter the inevitable result, efficiency of convergence is often the driving force in selecting one method over another. Many have found that hybrid methods combing the Metropolis-Hastings algorithm and Gibbs sampling emerge as the most efficient [Tierney, 1994]. We follow in these footsteps by employing the Metropolis-within-Gibbs algorithm [Metropolis et al., 1953, Tierney, 1994]. In particular, we will use this approach to draw inference on our covariates given $\mathcal{G}$.

In the Metropolis-within-Gibbs algorithm, we nest a Metropolis-Hastings algorithm within a Gibbs sampler, learning about $\boldsymbol{\beta}$ and σ^2 through two dependent processes. We use the Metroplis-Hastings algorithm to learn about $P(\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{X},\sigma^2)$. To implement the Metropolis-Hastings portion, we use an independence sampler for $\boldsymbol{\beta}$ with a normal transition kernel. Given a mode estimate $\hat{\boldsymbol{\beta}}$, we make a tentative draw of

$$\boldsymbol{\beta}^* \sim N(\hat{\boldsymbol{\beta}}, \frac{1}{\tau}\hat{\boldsymbol{\Sigma}}), \tag{5.9}$$

for positive τ. Following the standard form of the Metropolis-Hastings step, we compute

$$r = \frac{P(\boldsymbol{\beta}^*|\mathcal{G},\boldsymbol{Y},\boldsymbol{X},\sigma^2)}{P(\hat{\boldsymbol{\beta}}|\mathcal{G},\boldsymbol{Y},\boldsymbol{X},\sigma^2)}. \tag{5.10}$$

To accept or reject this proposal, we define

$$\delta = min(r, 1) \tag{5.11}$$

and sample

$$u \sim Unif(0, 1). \tag{5.12}$$

$\boldsymbol{\beta_t}$ is accepted if δ is greater than a sampled u.

Mixing remains a concern. Adaptive Metropolis-Hastings algorithms are designed to help with mixing problems [Roberts et al., 1997, Roberts and Rosenthal, 2009]. We address this problem in our model by adaptively selecting a tuning parameter τ that scales the variance of the proposal distribution kernel. Following Roberts et al. [1997], we strive for an acceptance frequency α around 0.25. Specifically, for MCMC iterate m, we tune τ as

$$\tau_m = \tau_{m-1} + \frac{1}{1+\sqrt{m}}(\alpha_i - \alpha). \tag{5.13}$$

We learn about the hyperprior σ^2 with the Gibbs sampler. Specifically, we have modeled our hyperprior with an inverse Gamma distribution. Therefore, we will sample the precision $\frac{1}{\sigma^2}$ from a $Gamma(\kappa, \theta)$ distribution. We define μ as the mean of $\hat{\boldsymbol{\beta}}$. For fixed constant κ_0 and θ_0, we draw

$$\frac{1}{\sigma^2} \sim Gamma(\kappa_0 + \frac{N}{2}, \theta_0 + \frac{1}{2}\sum(\hat{\beta}_{i-1} - \mu)^2). \tag{5.14}$$

After drawing $\frac{1}{\sigma^2}$, we recompute $\hat{\boldsymbol{\beta}}$ using cyclic coordinate descent.

5.1.3 Demonstration

5.1.3.1 Synthetic study: small illustration

We look to validate our model with a small synthetic dataset that illustrates the effectiveness of our approach. We simulate 1,000 patients exposed to 10 medical products. Among these products, 7 are safe, with log relative risks of 0. The other 3 products pose a risk to the simulated patients, with log relative risks of 0.2. We structure the simulated data in this way to underscore why model selection makes sense for comparative effectiveness and drug safety surveillance studies. Most medical products have no effect on a given outcome of interest. The true model of interest is therefore considerably smaller than the full model.

$$\boldsymbol{\beta}_{truth} = (0.2, 0, 0.2, 0, 0, 0, 0, 0, 0, 0.2)' \tag{5.15}$$

We first find that the model selection chooses the most reasonable models given our simulation framework. Showing the top 8 models, we see that model selection chooses, ordered by posterior density,

$$\boldsymbol{\beta}_{\mathcal{G}} = \begin{cases} \boldsymbol{\beta}_{g_1} = (\beta_0, 0, \beta_2, 0, 0, 0, 0, 0, 0, \beta_9)' \\ \boldsymbol{\beta}_{g_2} = (\beta_0, 0, \beta_2, 0, 0, 0, 0, \beta_7, 0, \beta_9)' \\ \boldsymbol{\beta}_{g_3} = (\beta_0, \beta_1, \beta_2, 0, 0, 0, 0, 0, 0, \beta_9)' \\ \boldsymbol{\beta}_{g_4} = (\beta_0, 0, \beta_2, 0, 0, \beta_5, 0, 0, 0, \beta_9)' \\ \boldsymbol{\beta}_{g_5} = (\beta_0, 0, \beta_2, 0, \beta_4, 0, 0, 0, 0, \beta_9)' \\ \boldsymbol{\beta}_{g_6} = (\beta_0, 0, \beta_2, 0, 0, 0, 0, 0, \beta_8, \beta_9)' \\ \boldsymbol{\beta}_{g_7} = (\beta_0, 0, \beta_2, 0, 0, 0, \beta_6, 0, 0, \beta_9)' \\ \boldsymbol{\beta}_{g_8} = (\beta_0, 0, \beta_9, \beta_3, 0, 0, 0, 0, 0, \beta_9)' \\ \dots \end{cases} . \tag{5.16}$$

Among the models shown, 92% of the density is placed on g_1, the true model, and the trail of models with less density are a single covariate inclusion away from the true model, as seen in Figure (5.1). Notably, among the top 8 models, none have dropped the 3 covariates that have true non-zero log relative risk.

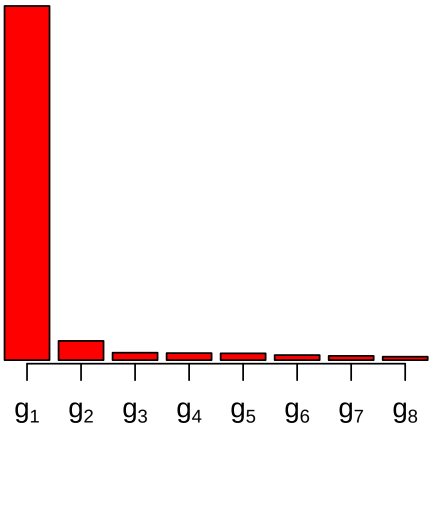


Figure 5.1: *The relative posterior density placed on the top 8 models selected using our Bayesian model selection approach.*

Working with these smaller models facilitates the Metropolis-within-Gibbs approach. To compare acceptance rates of our Metropolis-Hastings transition kernel, we fix $\tau = 1$. Under g_{full}, even in this low dimensional setting, the acceptance rate is 46%. As a comparison, under g_1 the acceptance rate is 75%. In the context of this toy synthetic data, these differences are irrelevant for overall convergence. However, as the dimensionality of the problem increases, the marginal benefit of using the model selection approach before the Metropolis-within-Gibbs simulations offers greater potential merit.

5.1.3.2 Small real world study: bleeding events

We also test our approach on a small, real world example. Using the small dataset introduced in Chapter 2, we revisit the problem of comparative risk for dangerous bleeding events between warfarin or dabigatran etexilate. In particular, we again examine the risk of gastrointestinal hemorrhage (GIH) and intracranial hemorrhage (IcH). We do not use our hierarchical model, but rather draw risk estimates ignoring the shared pathology. To perform these studies, we again examine the MarketScan Lab Results (MSLR) dataset, maintained by the Reagan-Udall Foundation Innovation in Medical Evidence Development and Surveillance project. Using the OMOP common data model version 4 for representation of concepts of interest, we collect all patients who experienced a diagnosis of IcH or GIH. There are 37,909 patients who had GIH and 2,893 patients who had IcH.

We can compare the results from Figure (2.2) and Figure (5.2). Qualitatively, a few trends are striking. First, both the bootstrap confidence intervals and the marginal posterior distributions reflect the same risk patterns, namely, warfarin shows the higher risk for IcH and the lower risk for GIH, relative to dabigatran. However, the differences between the outcome-specific risk distributions for warfarin are much more striking when looking at the marginal posterior distributions.

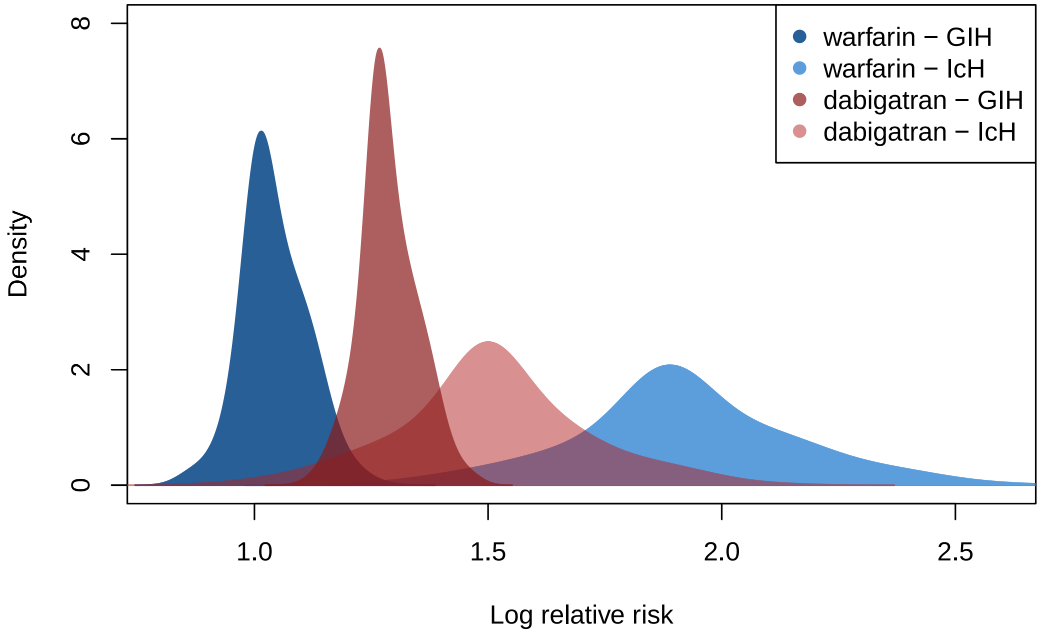


Figure 5.2: *Marginal densities for the relative risk of dabigatran and warfarin for gastrointestinal hemorrhage (GIH) and intracranial hemorrhage (IcH) using the MSLR dataset.*

5.1.4 Discussion

In this work, we leverage Bayesian model averaging to implement fully Bayesian inference at the scale of observational healthcare data. We accomplish this by averaging over covariate inclusion models. Then, given our model probabilities, we learn the relative risks for each included covariate as well as the hyperprior variance. Specifically, we rely on adaptive Metropolis-within-Gibbs.

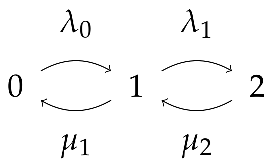
There are many opportunities for improvement in this project. First, we fail to include the graphic processing unit (GPU) mode finding strategy used by Suchard et al. [2013]. For the model averaging component of the project, mode finding remains the computational bottleneck. Therefore, using the GPU implementation will have a significant effect on run time. Furthermore, there are other opportunities for GPU parallelization in this project. In the mode finding approach of Suchard et al. [2013], the log likelihood undergoes rank one updates, as resetting all of the regression coefficients and recomputing the log likelihood does not occur. However, in the Metropolis-Hastings method we use, we frequently reset all of the $\boldsymbol{\beta}$ and recompute the log likelihood from scratch. The computationally expensive component to this step is a sparse matrix vector multiplication. We can implement this operation on the GPU as well.

5.2 Non-parametric treatment intensification

Although modeling treatment intensification as a birth-death process helps to elucidate the trajectory of patients through the space of oral anti-hyperglycemic medication, the simple birth-death model that we use left much to be desired. First of all, we did not provide a framework for learning about parameters other than HbA1c. Comorbidities, concurrent treatments, and previous adverse events could all reasonably alter treatment trajectory. For example, patients with renal disease are often advised against taking metformin, otherwise the first-line treatment of choice. Similarly, patients who experience hypoglycemic events may avoid more aggressive treatment regimens. These clinically relevant questions are beyond the scope our current approach. In a effort to mitigate these shortcomings, we propose an approach to modify our birth-death model.

5.2.1 Including covariates

Beginning with notation as before, let j index the states of the birth-death process, the number of drugs a patient is taking. Let λ_j be the birth rate for moving from j to $j+1$ drugs. Similarly, let μ_j be the death rate of moving from j to $j-1$ drugs. To account for the edge conditions, define the maximum number of drugs taken concurrently in a dataset as J. We enforce $\lambda_J = 0$ and $\mu_0 = 0$.



Before we start by considering covariates in our model, we are going to introduce a toy dataset that we will use to illustrate how each of our methods will be put to use. The graph in Figure (5.3) shows 3 example patient trajectories. The time is in arbitrary units, and we disregard HbA1c status, focusing solely on the drug count. Under the constant model, the likelihood is

$$\begin{aligned}\mathcal{L} =& e^{-(\lambda_1+\mu_1)3}\mu_1 \\ & \times e^{-(\lambda_1+\mu_1)4}\lambda_1 \times e^{-(\lambda_2+\mu_2)9}\lambda_2 \\ & \times e^{-(\lambda_1+\mu_1)5}\lambda_1 \times e^{-(\lambda_2+\mu_2)1}\mu_2.\end{aligned} \tag{5.17}$$

This allows us to estimate $\hat{\lambda}_1 = \frac{1}{6}$.

We want the transition rates to be non-parametric, multiplicatively modulated by the covariates of interest. We begin by considering the Cox proportional hazards model. In this model, the hazard function $\lambda(t, \boldsymbol{x})$ is given by

$$\lambda(t, \boldsymbol{x}) = \lambda_0(t) e^{\boldsymbol{x}'\beta}. \tag{5.18}$$

Let the time to event T have density $f(t, \boldsymbol{x})$ and distribution $F(t, \boldsymbol{x})$. The

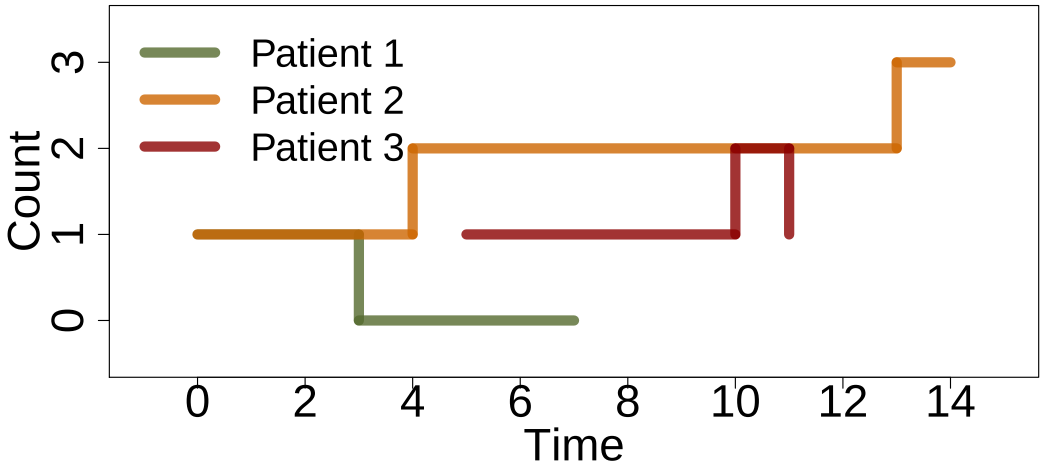


Figure 5.3: *This illustrates the treatment trajectories for 3 patients using oral anti-hyperglycemic medication. These data will be central to the methods discussion of integrating covariates.*

survival function $S(t, \boldsymbol{x})$ is

$$S(t, \boldsymbol{x}) = \frac{f(t, \boldsymbol{x})}{1 - F(t, \boldsymbol{x})} = \frac{f(t, \boldsymbol{x})}{S(t, \boldsymbol{x})}. \tag{5.19}$$

Equivalently,

$$S(t, \boldsymbol{x}) = e^{-\Lambda(t, \boldsymbol{x})} \tag{5.20}$$

where

$$\Lambda(t, \boldsymbol{x}) = \int_0^t \lambda(u, \boldsymbol{x}) du. \tag{5.21}$$

Let p index the patients. We start with a simple survival model. Consider $\boldsymbol{T} = \{t_1, t_2, \ldots t_P\}$ the set of observed event times, one per patient. Let δ_p indicate if t_p is censored or not. $\delta_p = 0$ if t_p is a censored event. We only consider right censoring.

The full likelihood under this model is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \prod_p [f(t_p, \boldsymbol{x}_p)]^{\delta_p} [S(t_p, \boldsymbol{x}_p)]^{1-\delta_p} \\ &= \prod_p [\lambda(t_p, \boldsymbol{x}_p)]^{\delta_p} [S(t_p, \boldsymbol{x}_p)] \end{aligned} \tag{5.22}$$

But this is cannot be optimized because $\lambda_0(t)$ is unknown. So we use the partial likelihood instead. To do this, we must introduce Q_{t_p}, the set of patients still observed at t_p. In the partial likelihood, we consider

$$\begin{aligned}\mathcal{L}_{partial}(\boldsymbol{\beta}) &= \prod_p \left[\frac{\lambda(t_p, \boldsymbol{x}_p)}{\sum_{q \in Q_{t_p}} \lambda(t_q, \boldsymbol{x}_q)}\right]^{\delta_p} \\ &= \prod_p \left[\frac{\lambda_0(t_p) e^{\boldsymbol{x}'_p \beta_p}}{\sum_{q \in Q_{t_p}} \lambda_0(t_q) e^{\boldsymbol{x}'_q \beta_q}}\right]^{\delta_p} \\ &= \prod_p \left[\frac{e^{\boldsymbol{x}'_p \beta_p}}{\sum_{q \in Q_{t_p}} e^{\boldsymbol{x}'_q \beta_q}}\right]^{\delta_p}\end{aligned} \tag{5.23}$$

where the final step is made possible by canceling the underlying non-parametric baseline hazard.

Two parts of the birth-death model are, in isolation, survival problems. In particular, birth from the first state and death from the last state can be modeled with the Cox proportional hazards model. Treating these steps independently, for constant λ_0 we substitute $\lambda_{0,0}(t_p)e^{\boldsymbol{x}'_p \beta_1}$. Similarly, for our simple model with up to 2 drugs, we substitute $\lambda_{2,0}(t_p)e^{\boldsymbol{x}'_p \beta_2}$ for μ_2. Others have recognized that the first step is a Cox proportional hazards model [Berkowitz et al., 2014].

Extending this insight, each of the other transitions emerge as separate competing risks problems. That is, for each state other than the edge cases, birth or death from that state represent competing events. An extension of the Cox proportional hazards model exists for the competing risks framework. For a model with k competing risks (causes), it is possible to have a cause-specific hazard

$$\lambda_k(t, \boldsymbol{x}) = \lambda_{k,0}(t) e^{\boldsymbol{x}' \beta_k}. \tag{5.24}$$

Note that there are cause-specific covariates β_k as well.

In the multiple cause scenario, we define $S(t, \boldsymbol{x})$ as the probability of surviving all types of events up to t. By analogy, we have

$$S_k(t, \boldsymbol{x}) = e^{-\Lambda_k(t, \boldsymbol{x})} \tag{5.25}$$

where

$$\Lambda_k(t, \boldsymbol{x}) = \int_0^t \lambda_k(u, \boldsymbol{x}) du. \tag{5.26}$$

Denoting outcome k_p as the outcome for person p, the total likelihood here is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_p [\lambda_{k_p}(t_p, \boldsymbol{x}_p)]^{\delta_p} [S(t_p, \boldsymbol{x}_p)]. \tag{5.27}$$

Note that $S(t_i, x_i) = \prod_k S_k(t_i, \boldsymbol{x}_i)$. Thus,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \prod_p [\lambda_{k_p}(t_p, \boldsymbol{x}_p)]^{\delta_p} \prod_k S_k(t_p, \boldsymbol{x}_p) \\ &= \prod_p \prod_k [\lambda_k(t_p, \boldsymbol{x}_p)]^{\delta_{p,k}} S_k(t_p, \boldsymbol{x}_p) \end{aligned} \tag{5.28}$$

where $\delta_{p,k}$ is a patient-outcome specific indicator such that $\delta_{p,k} = 1$ if and only if patient p experienced cause k.

This allows us to split the likelihood by cause type. Furthermore, for each cause of interest, we treat the other causes as censored points. The partial likelihood becomes

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\delta}) &= \prod_k \prod_p \frac{[\lambda_{k_p}(t_p, \boldsymbol{x}_p)]^{\delta_{p,k}}}{\sum_{q \in Q_{t_{k,p}}} [\lambda_{k_q}(t_q, \boldsymbol{x}_q)]^{\delta_{q,k}}} \\ &= \prod_k \prod_p \frac{[\lambda_{k_p,0}(t_p, \boldsymbol{x}_p) e^{\boldsymbol{x}'_p \beta_k}]^{\delta_{p,k}}}{\sum_{q \in Q_{t_{k,p}}} [\lambda_{k_q,0}(t_q, \boldsymbol{x}_q) e^{\boldsymbol{x}'_q \beta_k}]^{\delta_{q,k}}} \\ &= \prod_k \prod_p \left[\frac{e^{\boldsymbol{x}'_p \beta_k}}{\sum_{q \in Q_{t_{k,p}}} [e^{\boldsymbol{x}'_q \beta_k}} \right]^{\delta_{p,k}} \end{aligned} \tag{5.29}$$

5.2.2 Non-parametric birth-death process

Having identified subproblems within our birth-death process formulation that are amenable to non-parametric regression, we must still splice all of these components together together. There are a few obstacles to this goal. First, how we deal with time requires more consideration than in a single survival or competing risk model. If we take the naive approach and use some metric

$\Delta t_{k,m}$ the amount of time spent at state k in observation m, our baseline hazard becomes a $\lambda_{k,0}(\Delta t_{k,m})$. We loose the time dependent form that makes this model desirable in the first place. Second, one of the assumptions underlying the Cox proportional hazards model is the independence of observed survival events. Using the offset per state approach, many observations will not be independent, as patients will re-enter states and thus be present multiple times in the analysis.

We solve these issues by left truncating. This technique is commonly used to address staggered entrance into survival studies and is the preferable method for accounting for different patient ages. For observation m, let the drug count transition be at t_m. Furthermore, let v_m be the time at which observation m began. In the Cox proportional hazards partial likelihood, we consider the set of observations, Q_{t_m}, against which to compare observation m. Without left truncating,

$$Q_{t_m} = \{j : t_j > t_m\}. \tag{5.30}$$

With left truncating, we redefine

$$Q_{t_m} = \{j : v_j < t_m < t_j\}. \tag{5.31}$$

In other words, for each observation, we only compare it to the other observations whose time intervals contain t_m. Looking at our toy data set in Figure (5.3), we can recognize the transitions that can be evaluated and what their comparator sets contain. The first transition that we include in our model is the deescalation event of patient 1 from 1 drug to 0 drugs at time 3. The comparator set for this event includes both patient 1's own transition (an event) and patient 2's escalation to 2 drugs, which enters as a censored event. Notably, patient 3 also makes the transition from 1 drug to 2 drugs, but this would not enter into the comparator set because patient 3 is not extant at time 3. Patient 1 exists during patient 2's transition from 1 to 2 drugs at time 4, but they do not share a common state. Therefore, the transition from 1 to 2 for patient 2 does not enter into our analysis. In fact, the only two transitions that enter into the likelihood using left truncation from the toy dataset are the transition from 1 to 0 in patient 1 and the transition from 2 to 1 in patient 3.

Left truncation solves both problems listed above. First, we are no longer using the offset time from arrival to each state. Therefore, the model treats time universally. This allows the baseline hazard rate functions to be meaningfully interpreted. Second, a single patient cannot be at the same state during overlapping times. Therefore, we avoid the problem of dependence among the observations within each patient.

5.2.3 Dimensionality

Using our large observational datasets poses a dimensionality problem. Specifically, we need to draw inference on the effects estimates for all covariates of interest for each transition. That is, while $\boldsymbol{\beta}_{k,k+1}$, the regression covariates for transitioning from k to $k+1$, is large, $\boldsymbol{\beta} = [\boldsymbol{\beta}_{0,1}, \ldots, \boldsymbol{\beta}_{K-1,K}, \boldsymbol{\beta}_{K,K-1}, \ldots, \boldsymbol{\beta}_{1,0}$ is much larger. However, we have already developed a framework for dealing with problems like this in Chapter 2. We can consider using the hierarchical prior. This is medically reasonable because similar covariates should similarly affect time spent in each state.

Our sample size was notably reduced with left truncation in the toy dataset, as we went from 5 included transitions to 2. This will certainly be problematic for small datasets. Our large datasets come to our rescue. Because of our conditioning arguments, the number of patients in any given denominator will be small relative to the total number of patients in the dataset. With too little data, the number of overlapping intervals may be too small to perform any analysis. Therefore, we may only be able to fit this model because we are working in a high dimensional setting. This is a somewhat unique approach in that it succeeds only in the setting of massive amounts of data. We are empowered to use left truncation, the more rigorous approach, strictly because of the setting in which we are drawing inference.

Bibliography

GC Alexander, NL Sehgal, RM Moloney, and R Stafford. National trends in treatment of type 2 diabetes mellitus. *Arch Intern Med*, **168(19)**: 2088-94., 2008.

American Diabetes Association. Approaches to glycemic treatment. sec. 7. in standards of medical care in diabetes - 2015. *Diabetes Care*, **38(Suppl. 1)**: S41-S48., 2015.

P Barieri and M Maistrello. Usefulness of administrative databas- es for epidemiological evaluations and healthcare planning. *S.Co.2009 Politecnico di Milano*, 2009.

N Becker. Vaccination programs for rare infectious diseases. *Biometrika*, **59(2)**: 443-453., 1972.

CL Bennett, JR Nebeker, PR Yarnold, CC Tigue, DA Dorr, JM McKoy, BJ Edwards, JF Hurdle, DP West, DT Lau, C Angelotta, SA Weitzman, SM Belknap, B Djulbegovic, MS Tallman, TM Kuzel, AB Benson, A Evens, SM Trifilio, DM Courtney, and DW Raisch. Evaluation of serious adverse drug reactions: a proactive pharmacovigilance program (radar) vs safety activities conducted by the food and drug administration and pharmaceutical manufacturers. *Arch Intern Med*, **167(10)**: 1041-9., 2007.

SA Berkowitz, AA Krumme, J Avorn, T Brennan, OS Matlin, CM Spettell, DJ Pezalla, G Brill, WH Shrank, and NK Choudhry. Initial choice of oral glucose-lowering medication for diabetes mellitus a patient-centered comparative effectiveness study. *JAMA Intern Med*, **174(12)**: 1955-62., 2014.

SJ Boccuzzi, J Wogen, J Fox, JC Sung, AB Shah, and J Kim. Utilization of oral hypoglycemic agents in a drug-insured u.s. population. *Diabetes Care*, **24(8)**: 1411-5., 2001.

CP Cannon and PJ Cannon. COX-2 inhibitors and cardiovascular risk. *Science*, **336(6087)**: 1386–1387, 2012.

B Charlton and R Redberg. The trouble with dabigatran. *BMJ*, **349**: g4681, 2014.

A Chatterjee and SN Lahiri. Asymptotic properties of the residualbootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, **138**: 4497-4509., 2010.

A Chatterjee and SN Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, **106(494)**: 608-625., 2011.

SJ Connolly, MD Ezekowitz, Yusuf Salim, J Eikelboom, J Oldgren, A Parekh, J Pogue, PA Reilly, E Themeles, J Varrone, S Wang, M Alings, D Xavier, J Zhu, R Diaz, BS Lewis, H Darius, HC Diener, CD Joyner, L Wallentin, and the RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. *NEJM*, **361,12**: 228–35, 2009.

SJ Connolly, MD Ezekowitz, S Yusuf, PA Reilly, and L Wallentin. Newly identified events in the re-ly trial. *NEJM*, **363**: 1875-6, 2010.

AG Crawford, C Cote, J Couto, M Daskiran, C Gunnarsson, K Haas, SC Nigam, R Schuette, and J Yaskin. Comparison of ge centricity and electronic medical record database and national ambulatory medical care survey findings on the prevalence of major conditions in the united states. *Population Health Management*, **13(3)**: 139-150., 2010.

CJ Crooks, D Prieto-Merino, and SJW Evans. Identifying adverse events of vaccines using a Bayesian method of medically guided information sharing. *Drug Safety*, **35**: 61-78, 2012.

JR Curtis, H Cheng, E Delzell, D Fram, M Kilgore, K Saag, H Yun, and W Dumouchel. Adaptation of Bayesian data mining algorithms to longitudinal claims data: coxib safety as an example. *Medical Care*, **46(9)**: 969–75, 2008.

G Dailey, MS Kim, and JF Lian. Patient compliance and persistence with anti-hyperglycemic therapy: evaluation of a population of type 2 diabetic patients. *J Int Med Res*, **30(1)**: 71-79., 2002.

AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B.*, **39(1)**: 1 - 38, 1977.

W DuMouchel. Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, **53(3)**: 177Ð190., 1999.

W DuMouchel. Multivariate Bayesian logistic regression for analysis of clinical study safety issues. *Statistical Science*, **27**: 319-339, 2012.

B Efron and G Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, **37(1)**: 36-48., 1983.

C Farrington. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*, **51**: 228–35, 1995.

BR Flay, A Biglan, RF Boruch, FG Castro, D Gottfredson, S Kellam, EK Moscicki, S Schinke, JC Valentine, and P Ji. Standards of evidence: criteria for efficacy, effectiveness and dissemination. *Prevention Science: the official journal of the Society for Prevention Research*, **6(3)**: 151-175., 2005.

J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear modes via coordinate descent. *Journal of Statistical Software*, **33**: 1-22, 2010.

A Genkin, DD Lewis, and D Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, **49, 3**: 291-304, 2007.

EI George and RE McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**: 881-889., 1993.

M Gonzalez, P Patrignani, S Tacconelli, and LA Garcia Rodriguez. Variability among nonsteroidal antiinflammatory drugs in risk of upper gastrointestinal bleeding. *Arthritis Rheumatology*, **62(6)**: 1592-601, 2010.

M Grabner, Y Chen, M Nguyen, SD Abbott, and R Quimbo. Using observational data to inform the design of a prospective effectiveness study for a novel insulin delivery device. *ClinicoEconomics and Outcomes Research*, **5**: 471-479., 2013.

RT Grimes, K Bennett, L Tilson, C Usher, SM Smith, and MC Henman. Initial therapy, persistence and regimen change in a cohort of newly treated type

2 diabetes patients. *British Journal of Clinical Pharmacology*, **79(6)**: 1000-1009., 2015.

C Hampp, V Borders-Hemphill, DG Moeny, and DK Wysowski. Use of antidiabetic drugs in the u.s., 2003-2012. *Diabetes Care*, **37(5)**: 1367-74., 2014.

Y Handelsman, ZT Bloomgarden, G Grunberger, G Umpierrez, RS Zimmerman, TS Bailey, Blonde, GA Bray, AJ Cohen, S Dagogo-Jack, JA Davidson, D Einhorn OP Ganda, AJ Garber, WT Garvey, RR Henry, IB Hirsch, ES Horton, DL Hurley, PS Jellinger, L Jovanovic, HE Lebovitz, D LeRoith, P Levy, JB McGill, JI Mechanick, JH Mestman, ES Moghissi, EA Orzeck, R Pessah-Pollack, PD Rosenblit, AI Vinik, K Wyne, and F Zangeneh. American association of clinical endocrinologist and american college of endocrinology - clinical practice guidelines for developing a diabetes mellitus comprehensive care plan - 2015. *Endocr Pract.*, **21(4)**: 438-447., 2015.

MI Harris, RC Eastman, CC Cowie, KM Flegal, and MS Eberhardt. Racial and ethnic differences in glycemic control of adults with type 2 diabetes. *Diabetes Care*, **22**: 403-408., 1999.

AG Hartzema, CG Reich, PB Ryan, PE Stang, D Madigan, E Welebob, and JM Overhage. Managing data quality for a drug safety surveillance system. *Drug Safety*, **36(Suppl 1)**: S49-S58., 2013.

M Hauben, D Madigan, CM Gerrits, L Walsh, and EP Van Puijenbroek. The role of data mining in pharmacovigilance. *Expert Opinion on Drug Safety*, **4(5)**: 929-948, 2005.

C Hawkey, A Kahan, K Steinbr§ck, C Alegre, E Baumelou, B BŐgaud, J Dequeker, H IsomŁki, G Littlejohn, J Mau, and S Papazoglou. Gastrointestinal tolerability of meloxicam compared to diclofenac in osteoarthritis patients. international MELISSA study group. meloxicam large-scale international study safety assessment. *British Journal of Rheumatology*, **37(9)**: 937-45, 1998.

L Hazel-Fernandez, Y Xu, C Moretz, Y Meah, J Baltz, J Lian, E Kimball, and J Bouchard. Historical cohort analysis of treatment patterns for patients with type 2 diabetes initiating metformin monotherapy. *Curr Med Res Opin*, **31(9)**: 1703-1716., 2015.

G Hripcsak, P Ryan, J Duke, NH Shah, RW Park, V Huser, MA Suchard, M Schuemie, F DeFalco, A Perotte, J Banda, C Reich, L Schilling, M Matheny, D Meeker, N Pratt, and D Madigan. Addressing clinical questions at scale: Ohdsi assessment of treatment pathways. *Publication submitted*.

Y Huang, J Moon, and JB Segal. Comparison of active adverse event surveillance systems worldwide. *Drug Safety*, **37(8)**: 581Ð596., 2014.

Z Huang, X Lu, and H Duan. Latent treatment pattern discovery for clinical processes. *Journal of Medical Systems*, 37(2):1–10, 2013a.

Z Huang, X Lu, H Duan, and W Fan. Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, 46(1):111–127, 2013b.

DR Hunter and K Lange. Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, **9(1)**: 60 - 77, 2000.

DR Hunter and K Lange. A tutorial on MM algorithms. *American Statistician*, **58**: 30 - 37, 2004.

K Imai, G King, and EA Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171(2)**: 481-502., 2008.

V Irvine, S McClean, and P Millard. Stochastic models for geriatric in-patient behavior. *IMA Journal of Mathematics Applied in Medicine and Biology*, **11**: 207-216., 1994.

H Ishwaran and JS Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Journal of the American Statistical Association*, **33(2)**: 730-773., 2005.

DL Jaquette. A stochastic model for the optimal control of epidemics and pest populations. *Mathematical Biosciences*, **8(3-4)**: 343-354., 1970.

H Jick, GL Zornberg, SS Jick, S Seshadri, and DA Drachman. Statins and the risk of dementia. *Lancet*, **356**: 1627Ð1631., 2000.

N Keiding. Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics*, **3(2)**: 363-72., 1975.

DG Kendall. On the generalized “birth-and-death” process. *The Annals of Mathematical Statistics*, **19(1)**: 1-15., 1948.

M Kyung, J Gill, M Ghosh, and G Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, **5(2)**: 369Ð412, 2010.

A Lanas. A review of the gastrointestinal safety dataÑa gastroenterologistÕs perspective. *Rheumatology*, **49**: ii2-ii10, 2010.

K Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society. Series B*, **57(2)**: 425 - 437, 1995.

K Lange and TT Wu. An MM algorithm for multicategory vertex discriminant analysis. *Journal of Computational and Graphical Statistics*, **17(3)**: 527 - 544, 2008.

K Lange, DR Hunter, and I Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, **9**: 1 - 59, 2000.

DB Lawrence, KR Ragucci, LB Long, BS Parris, and LA Helfer. Relationship of oral antihyperglycemic (sulfonylurea or metformin) medication adherence and hemoglobin a1c goal attainment for hmo patients enrolled in a diabetes disease management program. *J Manag Care Pharm*, **12(6)**: 466-471., 1970.

JR Maclean, RH Chapman, CP Ferrufino, and G Krishnarajah. Drug titration patterns and hba 1c levels in type 2 diabetes. *Int J Clin Pract*, **63(7)**: 1008-1016., 2004.

M Maclure. The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, **133**: 144-53., 1991.

D Madigan, P Ryan, S Simpson, and Ivan Zorych. Bayesian methods in pharmacovigilance. In JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West, editors, *Bayesian Statistics 9*. Oxford University Press; Oxford, England: 2011, 2011.

D Madigan, PB Ryan, M Schuemie, PE Stang, JM Overhage, AG Hartzema, MA Suchard, W DuMouchel, and JA Berlin. Evaluating the impact of database heterogeneity on observational study results. *American Journal of Epidemiology*, **178**(4): 645-51., 2013.

D Madigan, PE Stand, JA Berlin, M Schuemie, M Overhage, MA Suchard, B Dumouchel, AG Hartzema, and PB Ryan. A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, **1** 11-39., 2014.

A Majeed, HG Hwang, SJ Connolly, JW Eidleboom, MD Exekowitz, L Wallentin, M Brueckmann, M Fraessdorf, S Yusuf, and S Schulman. Management and outcomes of major bleeding during treatment with dabigatran or warfarin. *Circulation*, **128**: 2325-2332, 2013.

N Metropolis, A Rosenbluth, M Rosenbluth, A Teller, and E Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**: 1087-1091., 1953.

TJ Mitchell and JJ Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, **83(404)**: 608-625., 1988.

D Nathan. Initial management of glycemia in type 2 diabetes mellitus. *New England Journal of Medicine*, **347**: 1342-1349., 2002.

J Overhage, P Ryan, C Reich, A Hartzema, and P Stang. Validation of a common data model for active safety surveillance research. *Journal of American Medical Informatics Association*, **19**: 54-60., 2012.

LS Phillips, WT Branch, CB Cook, JP Doyle, IM El-Kebbi, DL Gallina, CD Miller, DC Ziemer, and CS Barnes. Clinical inertia. *Annals of Internal Medicine*, **135**: 825-834., 2001.

M Pladevall, LK Williams, LA Potts, G Divine, H Xi, and JE Lafata. Clinical outcomes and adherence to medications measured by claims data in patients with diabetes. *Diabetes Care*, **27(12)**: 2800-2812., 2004.

E Poluzzi, E Raschi, U Moretti, and F De Ponti. Drug-induced tor- sades de pointes: data mining of the public version of the fda adverse event reporting system (aers). *Pharmacoepidemiol Drug Saf.*, **18(6)**: 512-518., 2009.

Y Qui, A Fu, M Davies, and S Engel. Underutilization of antihyperglycemic dual therapy in eligible, treatment naive patients with type 2 diabetes. *American Diabetes Association 72nd Scientific Sessions;*, 2012.

Y Qui, Q Li, J Tang, CPS Fan, Z Li, M Apecechea, R Hegar, R Shankar, KM Kurtyka, and SS Engel. Why physicians do not initiate dual therapy as recommended by aace guidelines: A survey of clinicians in the united states. *Diabetes Research and Clinical Practice;*, **108**: 456-465, 2015.

PA Reilly, T Lehr, S Haertter, SJ Connolly, S Yusuf, JW Eikelboom, MD Ezekowitz, G Nehmiz, S Wang, and L Wallentin on behalf of the RE-LY Investigators. The effects of dabigatran plasma concentrations and patient characteristics on the frequency of ischemic stroke and major bleeding in atrial fibrillation patients. *J Am Coll Cardiol*, **63**: 321-8, 2014.

SJ Reisinger, PB Ryan, DJ O'Hara, GE Powell, JL Painter, EN Pattishall, and JA Morris. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of American Medical Informatics Association*, **17(6)**: 652-62., 2010.

GO Roberts and JS Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, **18(2)**: 349-367., 2009.

GO Roberts, A Gelman, and WR Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, **7(1)**: 110-120., 1997.

JK Rothman, S Greenland, and T Lash. *Modern Epidemiology*. Wolters Kluwer, Philadelphia, PA, 3rd edition edition, 2008.

PB Ryan. Statistical challenges in systematic evidence generation through analysis of observational healthcare data networks. *Statistical Methods in Medical Research*, **22**(1):3-6, 2013.

S Schneeweiss. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiology and Drug Safety*, **19**: 858Ð868., 2014.

S Schneeweiss and J Avorn. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*, **58(4)**: 323Ð337., 2005.

J Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, **7**: 221-264., 1997.

J Shao and JS Rao. The GIC for model selection: a hypothesis testing approach. *Journal of Statistical Planning and Inference*, **88(2)**: 215Ð231., 2000.

PF Short, DR Graefe, and C Schoen. Churn, churn, churn: how instability of health insurance shapes America's uninsured problem. *Issue brief The Commonwealth Fund,*, New York, NY., 2003.

PF Short, DR Graefe, and C Schoen. Self-controlled methods for postmarketing drug safety surveillance in large-scale longitudinal data. *Dissertation*, Columbia University, NY., 2011.

SE Simpson, D Madigan, I Zorych, MJ Schuemie, PB Ryan, and MA Suchard. Multiple self-controlled cases series for large-scale longitudinal observational databases. *Biometrics*, **69**: 893–902, 2013.

SL Slabaugh, Y Xu, JN Stacy, JC Baltz, YA Meah, J Lian, DC Moretz, and JR Bouchard. Antidiabetic treatment patterns in a medicare advantage population in the united states. *Drugs Aging.*, **27(12)**: 169-78., 2015.

PE Stang, PB Ryan, JA Racoosin, JM Overhage, AG Hartzema, C Reich, E Welebob, T Scarnecchia, and J Woodcock. Advancing the science for active surveillance: Rationale and design for the observational medical outcomes partnership. *Annals of Internal Medicine*, **153(9)**: 600–6., 2010.

MA Suchard, SE Simpson, I Zorych, P Ryan, and D Madigan. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Transactions on Modeling and Computer Simulation*, **23(1)**: 1–17, 2013.

B Thuraisingham, L Khan, M Awad, and L Wang. *Design and Implementation of Data Mining Tools*. CRC Press, 2009.

RJ Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58(1)**: 267-288., 1996.

RJ Tibshirani. Least angle regression. *The Annals of Statistics*, **32(2)**: 407-499., 2004.

L Tierney. Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, **22(4)**: 1701-1762., 1994.

L Tierney and JB Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81(393)**: 82Ð86., 1986.

RC Turner, CA Cull, V Frighi, and RR Holman. Glycemic control with diet, sulfonylurea, metformin, or insulin in patients with type 2 diabetes mellitus progressive requirement for multiple therapies (ukpds 49). *JAMA*, **281(21)**: 2005-2012., 1999.

RA Vigersky, K Fitzner, and J Levinson. Barriers and potential solutions to providing optimal guideline-driven care to patients with diabetes in the u.s. *Diabetes Care*, **36**: 3843Ð3849., 2013.

W Weng, Y Liang, ES Kimball, T Hobbs, S Kong, B Sakurada, and J Bouchard. Drug usage patterns and treatment costs in newly-diagnosed type 2 diabetes mellitus cases, 2007 vs 2012: findings from a large us healthcare claims database analysis. *J Med Econ*, **26**: 1-8., 2016.

TT Wu and K Lange. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, **2(1)**: 224-244, 2008.

TT Wu, Y Chen, T Hastie, E Sobel, and K Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25(6)**: 714Ð721, 2009a.

TT Wu, YF Chen, T Hastie, E Sobel, and K Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25(6)**:714-21, 2009b.

D Yoon, I Park, MJ Schuemie, MY Park, JH Kim, and RW Park. A quantitative method for assessment of prescribing patterns using electronic health records. *PLoS ONE*, 8(10):e75214, 2013.

T Zhang and F Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, **4(1)**: 5Ð31, 2001.

H Zhou and Y Zhang. EM vs MM: A case study. *Computational Statistics and Data Analysis*, **56**: 3909 - 20, 2012.

H Zhou, K Lange, and MA Suchard. Graphics processing units and high-dimensional optimization. *Statistical Science*, **25(3)**: 311-324, 2010.

H Zhou, D Alexander, and K Lange. A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistical Computation*, **21(2)**: 261-273, 2011.