Pharmacovigilance via Baseline Regularization with Large-Scale Longitudinal Observational Data

Zhaobin Kuang
University of Wisconsin-Madison
zkuang@wisc.edu

Peggy Peissig
Marshfield Clinic
Peissig.Peggy@mcrf.mfldclin.edu

Vítor Santos Costa
Universidade do Porto
vsc@dcc.fc.up.pt

Richard Maclin
University of Minnesota-Duluth
rmaclin@d.umn.edu

David Page
University of Wisconsin-Madison
page@biostat.wisc.edu

ABSTRACT

Several prominent public health incidents [29] that occurred at the beginning of this century due to adverse drug events (ADEs) have raised international awareness of governments and industries about pharmacovigilance (PhV) [6, 7], the science and activities to monitor and prevent adverse events caused by pharmaceutical products after they are introduced to the market. A major data source for PhV is large-scale longitudinal observational databases (LODs) [6] such as electronic health records (EHRs) and medical insurance claim databases. Inspired by the Multiple Self-Controlled Case Series (MSCCS) model [27], arguably the leading method for ADE discovery from LODs, we propose baseline regularization, a regularized generalized linear model that leverages the diverse health profiles available in LODs across different *individuals* at different *times*. We apply the proposed method as well as MSCCS to the Marshfield Clinic EHR. Experimental results suggest that incorporating the heterogeneity among different patients and different times help to improve the performance in identifying *benchmark* ADEs from the Observational Medical Outcomes Partnership ground truth [26].

CCS CONCEPTS

•Mathematics of computing → Regression analysis; •Applied computing → Health care information systems; Health informatics;

KEYWORDS

Longitudinal Data; Electronic Health Records; Adverse Drug Event Discovery; Pharmacovigilance; Baseline Regularization

ACM Reference format:
Zhaobin Kuang, Peggy Peissig, Vítor Santos Costa, Richard Maclin, and David Page. 2017. Pharmacovigilance via Baseline Regularization with Large-Scale Longitudinal Observational Data. In *Proceedings of KDD'17, August 13–17, 2017, Halifax, NS, Canada., ,* 10 pages.
DOI: http://dx.doi.org/10.1145/3097983.3097998

KDD'17, August 13–17, 2017, Halifax, NS, Canada.
 978-1-4503-4887-4/17/08.
DOI: http://dx.doi.org/10.1145/3097983.3097998

1 INTRODUCTION

Pharmacovigilance (PhV) [6, 7] is the science and activities relating to the surveillance and prevention of adverse events caused by pharmaceutical products *after* they are introduced to the market. In response to several recent prominent public health hazards [29] due to adverse drug events (ADEs), governments, industries, and other stakeholders across the world have been building effective PhV systems to safeguard admissible profit-risk profiles of drug products on the market.

Major PhV systems [3, 8, 20] nowadays leverage a network of large-scale longitudinal observational databases (LODs) [6] such as electronic health records (EHRs) and medical insurance claim databases that contain individual-level time-stamped rich medical data collected globally from hundreds of millions of individuals. All the databases within the network are updated periodically and are converted to the same format; various ADE discovery algorithms can hence be run regularly on different databases without any modifications to achieve proactive drug safety surveillance.

An *efficient* algorithm that can deliver *accurate* ADE identification using LODs is hence of utmost importance to the performance of PhV systems. A leading algorithm is the Multiple Self-Controlled Case Series (MSCCS) method [27]. Using the occurrence of a condition of interest from different patients at different times as the response variable, and the corresponding exposure statuses of various drugs as the features, MSCCS is a parsimonious representation of a fixed effect Poisson regression model [34]. In MSCCS, each patient acts as his or her own control, during exposed (case) or unexposed (control) periods of time, thus controlling even for latent and unconsidered factors, provided they are *time-invariant*.

However, due to the longitudinal nature of the data, simply adjusting for time-invariant confounding does not suffice to deliver accurate modeling. For example, the occurrence rate of adverse events such as heart attacks usually increases as the observed individual ages. Moreover, patients that previously had heart attacks will also be prone to have another one in the future. Neither of the aforementioned *time-varying* occurrence rates of heart attack can be modeled by adjusting for time-invariant confounding via MSCCS.

By assuming an individual-specific, time-dependent occurrence rate of adverse events, the mission of the proposed Baseline Regularization (BR) method is to provide flexibility to model the temporal nature of LODs, in the hope of delivering more effective ADE discovery. Our contributions are three-fold:

- BR is the first general-purpose ADE discovery algorithm following a self-controlled design that exploits the time-varying perspective of individual profiles in large-scale LODs.
- BR is deeply connected to and is a generalization of some of the existing models in the literature. BR not only directly generalizes MSCCS, it is also a generalized linear model that extends [10], which deals with baseline regularization in a linear model setting.
- Experimental results suggest that incorporating the heterogeneity among different patients and different times help to improve the performance in identifying *benchmark* ADEs from the Observational Medical Outcomes Partnership ground truth [26].

2 MODEL SPECIFICATION

2.1 Background

Figure 1 visualizes the EHR from a patient that has taken two drugs and has had four heart attacks throughout his 400 days of observation. The rectangular bands in different colors represent different *drug eras*, each representing a consecutive time period during which the patient was exposed to a particular drug. A drug era is recorded with its *start date*, *end date*, and the name of the drug. The black arrows pointing downwards annotated with MI (Myocardial Infarction) represent the date on which the patient had a heart attack. The gray dashed lines and the indices on the top of the figure represent different *intervals*, a concept that we will define later in Section 2.3. In this paper, we consider the *multiple-drug, single-ADE* setting. As an illustrative example, our task of using the EHR from the patient presented in Figure 1 and from *many* other patients is to determine whether the exposure to certain drugs might cause the occurrence of MI as an adverse event.

Suppose there are M drugs and N patients in the EHR database. We use J_i to represent the total number of days of observation available in the EHR of patient i, where $i \in \{1, 2, \cdots, N\}$. We use χ_{ijm} to represent a binary drug exposure status of drug m on the j^{th} day during the observation of the i^{th} patient, where $j \in \{1, 2, \cdots, J_i\}$, and $m \in \{1, 2, \cdots, M\}$. $\chi_{ijm} = 1$ represents exposure and $\chi_{ijm} = 0$ represents non-exposure. We further use y_{ij} to represent a binary MI occurrence variable with $y_{ij} = 1$ meaning that the i^{th} patient has an MI on the j^{th} day during the observation, and $y_{ij} = 0$ otherwise. With the notation introduced above, we can consider y_{ij}'s as a response variable and χ_{ijm}'s as features. Following the convention of MSCCS, we will use a Poisson regression model (instead of a logistic regression model even though the response is binary) to depict the relationship between the response variable and the features, resulting in the following log-likelihood function:

$$\log \mathcal{L}(\boldsymbol{\tau}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{J_i} y_{ij} \left(\tau_{ij} + \boldsymbol{\chi}_{ij}^\top \boldsymbol{\beta} \right) - \exp\left(\tau_{ij} + \boldsymbol{\chi}_{ij}^\top \boldsymbol{\beta} \right), \quad (1)$$

where

$$\boldsymbol{\chi}_{ij} = \begin{bmatrix} \chi_{ij1} & \chi_{ij2} & \cdots & \chi_{ijM} \end{bmatrix}^\top, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_M \end{bmatrix}^\top,$$
$$\boldsymbol{\tau} = \begin{bmatrix} \tau_{11} & \tau_{12} & \cdots & \tau_{1J_1} & \cdots & \tau_{N1} & \tau_{N2} & \cdots & \tau_{NJ_N} \end{bmatrix}^\top.$$

The occurrence rate of MI to the i^{th} patient on the j^{th} day during observation is hence given by $\exp\left(\tau_{ij} + \boldsymbol{\chi}_{ij}^\top \boldsymbol{\beta} \right)$, from which we can infer that the rate is determined by two contributing factors. The first one depends on the joint drug exposure statuses, described by $\boldsymbol{\chi}_{ij}$, and the effect of each drug on the occurrence rate of MI, given by $\boldsymbol{\beta}$. If the value of a particular component of $\boldsymbol{\beta}$ is especially large, then the occurrence rate of MI will increase upon the exposure of the corresponding drug. Therefore, such a drug *might* potentially cause MI as an ADE. The second factor is the *baseline parameter* τ_{ij}, which models the inherent occurrence rate of MI for the i^{th} patient on day j excluding the interference of the effects from other covariates modeled by $\boldsymbol{\beta}$.

2.2 Baseline Regularization

Baseline Parameters

The introduction of the baseline parameters τ_{ij}'s in (1) is strikingly simple, and yet it offers tremendous flexibility to portray the heterogeneity of adverse event occurrence rates among different patients, and during different time periods within the same patient.

For example, a person who has High Blood Pressure (HBP) might have an inherently higher risk for heart attack compared with a healthy person. Therefore, the baseline parameters for the HBP patients might be higher compared with those of a healthy person. Within the same individual, commonsense-supported observations in the EHRs often suggest that one should vary baseline parameters temporally: for example, the risk for heart attack tends to increase in general as a person ages; a patient who has a history of heart attack might also be more likely to have another heart attack in the future. In both cases, a set of baseline parameters with increasing tendency along time within the same patient might be introduced to model such observations.

On the other hand, MSCCS makes the following more restrictive modeling assumptions:

$$\tau_{ij} = \alpha_i, \quad \forall i \in \{1, 2, \cdots, N\}, \quad \forall j \in \{1, 2, \cdots, J_i\}.$$

That is, MSCCS assumes that baseline parameters can only differ among different patients. Within the same patient, baseline parameters do not vary across time. While this modeling assumption is reasonable to address for time-invariant confounding such as gender, socioeconomic status, and genetic profile, it easily fails to model the aforementioned time-varying occurrence rates.

Regularization

An observant reader might have already noticed that the modeling flexibility introduced by baseline parameters τ_{ij}'s in (1) comes with the steep cost of *overparameterization*: the number of baseline parameters introduced is equal to the sample size of the data! Furthermore, in a typical EHR setting there could be thousands of drugs available. Modeling the effects of all these drugs will introduce a $\boldsymbol{\beta}$ whose dimension is easily on the order of thousands. The high dimensionality of both $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ motivates us to reduce the degrees of freedom of the model via *sparse* regularization, which results in the *baseline regularization* optimization problem as follows:

$$\arg\min_{\boldsymbol{\tau}, \boldsymbol{\beta}} -\log \mathcal{L}(\boldsymbol{\tau}, \boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \sum_{i=1}^{N} \sum_{j=1}^{J_i - 1} \lambda_2 |\tau_{i,j+1} - \tau_{ij}| + \lambda_3 \|\boldsymbol{\tau}\|_2^2. \quad (2)$$

Here in (2) we use a lasso penalty to regularize $\boldsymbol{\beta}$ because we assume that among thousands of drugs, there can only be a few

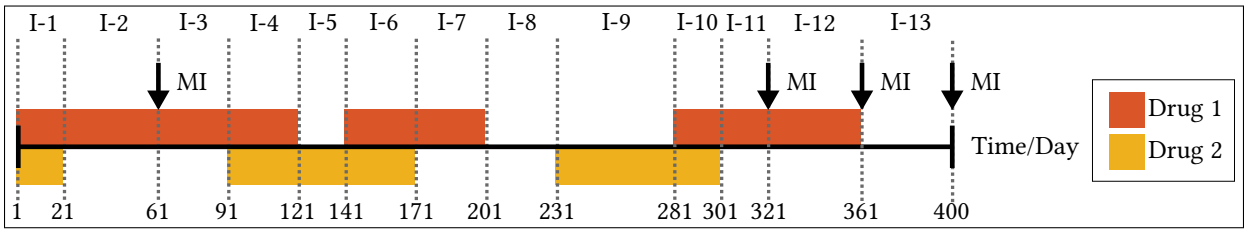


Figure 1: Visualization of a patient's EHR. MI: Myocardial Infarction (heart attack).

that influence the occurrence rate of MI. We use a fused lasso penalty [9, 19, 31] to regularize $\boldsymbol{\tau}$. The intuition behind using this penalty is that we assume the change between two *adjacent* baseline parameters is steady and gradual, and hence the baseline occurrence rate should not differ much from one day to another between two days that are adjacent to each other.

We also use a ridge penalty to regularize $\boldsymbol{\tau}$. The necessity for including this penalty can be seen from the observations between day 201 and day 230 in Figure 1. During this time period (interval I-8), $y_{ij} = 0$, and $\boldsymbol{\chi}_{ij} = \mathbf{0}$, $\forall j \in \{201, 201, \cdots, 230\}$, where for convenience we assume that the patient in Figure 1 is indexed by i. Therefore, during this time period, τ_{ij}'s will tend to be very negative in order to drive the occurrence rate $\exp(\tau_{ij})$ to a number that is very close to zero for a maximum likelihood interpretation of the data. In this scenario, a very negative τ_{ij} might overfit the data. Therefore, a ridge penalty that encourages smaller magnitudes of τ_{ij}'s is desirable to avoid overfitting. Furthermore, a ridge penalty can also encourage similarity among different components of $\boldsymbol{\tau}$, which also helps to foster the assumption that adjacent baseline occurrence rates should not differ much from one another. Using a ridge penalty is a common practice in many other densely parameterized machine learning models, with the most famous and popular example being (deep) neural networks [5, 13].

2.3 Scaling up Baseline Regularization

Even with the regularization introduced in (2), the computational burden of solving the BR model can still be staggeringly heavy. This is because a typical EHR database can easily contain billions of days of observations from all the patients; each day will require a separate baseline parameter to describe the baseline occurrence rate of an adverse event.

Intervals

To achieve scalability without much loss of modeling flexibility, we learn lessons from the idea of *data squashing* [11, 27] that exploits the discreteness and the sparsity of the data under consideration. Specifically, within the observational history of a particular patient, we define an *interval* as a consecutive time period during which the drug exposure statuses of *all* drugs *and* the cumulative number of adverse event occurrences remain unchanged.

Based on this definition, Figure 1 visualizes a patient's EHR that is divided into thirteen intervals. Each interval is indexed by I-k on the top of the figure, where $k \in \{1, 2, \cdots, 13\}$. The start date of each interval is passed through by a gray dashed line. Therefore, a previous interval ends right before a dashed line. For example, inclusively, I-1 starts from day 1 and ends at day 20 instead of day 21. Similarly, I-2 starts from day 21 and ends at day 60 instead of day 61. An exception for the unchanged cumulative adverse event occurrence restriction upon an interval is allowed if an adverse event occurs at the end of the observation. For example, in Figure 1, we consider I-13 ranges from day 361 to the end of the observation (day 400) even if on the last day there is a new occurrence of MI. The reason for allowing such an exception is to avoid a short (one day) interval at the end of an observation.

The concept of an interval provides convenience in describing the data concisely, and hence achieves the goal of data squashing. In Figure 1, instead of describing the data using information from 400 days, we can now use information from only thirteen intervals.

Parameter Tying

To reduce the number of baseline parameters used for modeling, we tie similar parameters together to the same value. Specifically, we consider two parameter tying strategies.

- **Interval Tying**: We can consider that the baseline parameters within the same interval are the same. In this case, within a patient, the number of baseline parameters used is equal to the number of intervals instead of the number of days of observation. In Figure 1, this parameter tying strategy reduces the number of baseline parameters from 400 to thirteen.
- **Occurrence Tying**: We can even further tie baseline parameters from similar intervals together. For example, since ADEs are usually recurrent, and the baseline risk of getting a subsequent ADE usually changes compared with getting the first one, we can tie intervals that have the same cumulative number of adverse event occurrences together. In Figure 1, this parameter tying strategy will further reduce the number of baseline parameters from thirteen to four, partitioned as:

$$\{\{\text{I-1}, \text{I-2}\}, \{\text{I-3}, \cdots, \text{I-11}\}, \{\text{I-12}\}, \{\text{I-13}\}\}.$$

Reformulation

We now reformulate the BR model in (2) using intervals and parameter tying. Let K_i denote the number of intervals that the EHR of the i^{th} patient is partitioned into. Let κ_i represent the number of baseline parameters used in BR after parameter tying either via interval tying or via occurrence tying. We define the vector of baseline parameters after tying as:

$$\boldsymbol{t} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1\kappa_1} & \cdots & t_{N1} & t_{N2} & \cdots & t_{N\kappa_N} \end{bmatrix}^\top,$$

Then the baseline parameter for each interval can also be represented as a vector: $\mathbf{Z}\boldsymbol{t}$, where $\mathbf{Z}$ is a $\left(\sum_{i=1}^{N} K_i\right) \times \left(\sum_{i=1}^{N} \kappa_i\right)$ binary

design matrix that maps the tied baseline parameters to the baseline parameters for each interval. Note that if the interval tying strategy is adopted, then $\kappa_i = K_i$, and $\mathbf{Z} = \mathbf{I}$, where $\mathbf{I}$ represents an identity matrix. This is because under the interval tying strategy, each interval will have its own baseline parameter. Furthermore, we use l_{ik} to represent the duration (time length) of the k^{th} interval from the i^{th} patient, where $k \in \{1, 2, \cdots, K_i\}$. And we use n_{ik} to represent the number of adverse event occurrences during the k^{th} interval of the i^{th} patient. We further use an $M \times 1$ binary vector $\boldsymbol{x}_{ik}$ to represent the drug exposure statuses during the k^{th} interval of the i^{th} patient. The reason why we need only one exposure vector to represent multiple days within an interval is due to the property of unchanged drug exposure statuses of any one interval. Stacking up l_{ik}'s, n_{ik}'s, and $\boldsymbol{x}_{ik}$'s results in their vector and matrix representations:

$$\begin{aligned}
\boldsymbol{l} &= \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1K_1} & \cdots & l_{N1} & l_{N2} & \cdots & l_{NK_N} \end{bmatrix}^\top, \\
\boldsymbol{n} &= \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1K_1} & \cdots & n_{N1} & n_{N2} & \cdots & n_{NK_N} \end{bmatrix}^\top, \\
\mathbf{X} &= \begin{bmatrix} \boldsymbol{x}_{11} & \boldsymbol{x}_{12} & \cdots & \boldsymbol{x}_{1K_1} & \cdots & \boldsymbol{x}_{N1} & \boldsymbol{x}_{N2} & \cdots & \boldsymbol{x}_{NK_N} \end{bmatrix}^\top.
\end{aligned}$$

Using $\mathbf{Z}$, $\boldsymbol{t}$, $\mathbf{X}$, $\boldsymbol{\beta}$, $\boldsymbol{l}$, and $\boldsymbol{n}$, we can rewrite the log-likelihood function in (1) in a matrix and vector form as follows:

$$\log \mathcal{L}(\boldsymbol{t}, \boldsymbol{\beta}) = \boldsymbol{n}^\top (\mathbf{Z}\boldsymbol{t} + \mathbf{X}\boldsymbol{\beta}) - \boldsymbol{l}^\top \exp(\mathbf{Z}\boldsymbol{t} + \mathbf{X}\boldsymbol{\beta}), \tag{3}$$

where $\exp(\cdot)$ represents a component-wise exponentiation.

After parameter tying, the fused lasso penalties imposed on $\boldsymbol{\tau}$ in (2) become fused lasso penalties imposed on the adjacent components of $\boldsymbol{t}$ that are from the same patient because under parameter tying:

$$\sum_{i=1}^{N} \sum_{j=1}^{J_i - 1} \lambda_2 |\tau_{i,j+1} - \tau_{ij}| = \sum_{i=1}^{N} \sum_{k=1}^{\kappa_i - 1} \lambda_2 |t_{i,k+1} - t_{ik}|. \tag{4}$$

We define $\mathbf{D}_q$ and $\mathbf{D}$ as follows:

$$\mathbf{D}_q = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \end{bmatrix}_{(q-1)\times q}, \quad \mathbf{D} = \begin{bmatrix} \mathbf{D}_{\kappa_1} & & & \\ & \mathbf{D}_{\kappa_2} & & \\ & & \ddots & \\ & & & \mathbf{D}_{\kappa_N} \end{bmatrix}, \tag{5}$$

where $\mathbf{D}_q$ is a $(q-1) \times q$ *first difference* matrix, and $\mathbf{D}$ is a *blockwise* first difference matrix. Note that $q \in \mathbb{N}_+$, and we define $D_1 = 0$.

With (2), (3), (4), and (5), we can reformulate the BR problem compactly as:

$$\arg\min_{\boldsymbol{t}, \boldsymbol{\beta}} -\log \mathcal{L}(\boldsymbol{t}, \boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\mathbf{D}\boldsymbol{t}\|_1 + \lambda_3 \|\boldsymbol{t}\|_2^2, \tag{6}$$

where we impose the same strength of ridge regularization using λ_3 on all the components of $\boldsymbol{t}$.

3 OPTIMIZATION ALGORITHM

This section provides an optimization algorithm for solving the compact BR model in (6). Following the idea of `glmnet` [4], we adopt an iteratively reweighted least squares (IRLS) approach to quadratically approximate the negative log-likelihood function. Observe that both the negative log-likelihood function and its quadratic approximation are convex, and $\boldsymbol{\beta}$ and $\boldsymbol{t}$ are *separable* in the regularization terms; we hence can perform blockwise minimization that alternates between $\boldsymbol{\beta}$ and $\boldsymbol{t}$ to achieve convergence [32].

3.1 Quadratic Approximation

At iteration p, the iterates $\boldsymbol{t}^{(p)}$ and $\boldsymbol{\beta}^{(p)}$ are given. We therefore can perform a quadratic approximation of (3) centered at the current iterates, in order to search for the next iterates that are closest to optimality in the vicinity of the current iterates. Optimizing the quadratic approximation is equivalent to solving a weighted least squares problem as follows:

$$\arg\min_{\boldsymbol{t}, \boldsymbol{\beta}} \frac{1}{2} \left\| \boldsymbol{z}^{(p)} - \mathbf{Z}\boldsymbol{t} - \mathbf{X}\boldsymbol{\beta} \right\|_{\mathbf{W}^{(p)}}^2, \tag{7}$$

where the working response is:

$$\boldsymbol{z}^{(p)} = \mathbf{Z}\boldsymbol{t}^{(p)} + \mathbf{X}\boldsymbol{\beta}^{(p)} + {\mathbf{W}^{(p)}}^{-1}\boldsymbol{n} - \mathbf{1}, \tag{8}$$

with $\mathbf{W}^{(p)} = \mathbf{L}\mathbf{S}^{(p)}$. $\mathbf{L} = \operatorname{diag} \boldsymbol{l}$, and $\mathbf{S}^{(p)} = \operatorname{diag} \boldsymbol{s}^{(p)}$ are diagonal matrices constructed by the elements of $\boldsymbol{l}$ and $\boldsymbol{s}^{(p)}$ respectively, with $\boldsymbol{s}^{(p)} = \exp\left(\mathbf{Z}\boldsymbol{t}^{(p)} + \mathbf{X}\boldsymbol{\beta}^{(p)}\right)$; $\mathbf{1}$ is a column vector of all ones, and $\|\boldsymbol{a}\|_{\mathbf{W}}^2 = \boldsymbol{a}^\top \mathbf{W} \boldsymbol{a}$, with $\boldsymbol{a}$ being a column vector and $\mathbf{W}$ being a positive diagonal matrix.

The derivation of the quadratic approximation for (3) basically follows from deriving the quadratic approximation of a standard Poisson regression model and the details are provided in the Appendix.

3.2 Blockwise Minimization

With quadratic approximation, at iteration p with the iterates $\boldsymbol{t}^{(p)}$ and $\boldsymbol{\beta}^{(p)}$ available, the next iterates can be obtained by considering the following optimization problem:

$$\begin{aligned}
\boldsymbol{t}^{(p+1)}, \boldsymbol{\beta}^{(p+1)} = \arg\min_{\boldsymbol{t}, \boldsymbol{\beta}} \frac{1}{2} \left\| \boldsymbol{z}^{(p)} - \mathbf{Z}\boldsymbol{t} - \mathbf{X}\boldsymbol{\beta} \right\|_{\mathbf{W}^{(p)}}^2 \\
+ \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\mathbf{D}\boldsymbol{t}\|_1 + \lambda_3 \|\boldsymbol{t}\|_2^2.
\end{aligned} \tag{9}$$

We will adopt a blockwise minimization strategy that fixes $\boldsymbol{t}$ and $\boldsymbol{\beta}$ alternatively and solves for the other until the iterates reach the optimality of (9). The optimization can hence be formulated as iterating between two steps: a $\boldsymbol{\beta}$-step and a $\boldsymbol{t}$-step.

$\boldsymbol{\beta}$-Step

We first initialize $\tilde{\boldsymbol{t}} = \boldsymbol{t}^{(p)}$. For each $\boldsymbol{\beta}$ step, we fix $\boldsymbol{t} = \tilde{\boldsymbol{t}}$ and solve the subproblem with respect to only $\boldsymbol{\beta}$ for $\tilde{\boldsymbol{\beta}}$:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \boldsymbol{z}^{(p)} - \mathbf{Z}\tilde{\boldsymbol{t}} - \mathbf{X}\boldsymbol{\beta} \right\|_{\mathbf{W}^{(p)}}^2 + \lambda_1 \|\boldsymbol{\beta}\|_1, \tag{10}$$

which is an L_1-regularized linear regression problem that can be solved efficiently by existing packages.

$\boldsymbol{t}$-Step

For each $\boldsymbol{t}$ step, we fix $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$, and solve the subproblem with respect to only $\boldsymbol{t}$ for $\tilde{\boldsymbol{t}}$:

$$\tilde{\boldsymbol{t}} = \arg\min_{\boldsymbol{t}} \frac{1}{2} \left\| \boldsymbol{z}^{(p)} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{t} \right\|_{\mathbf{W}^{(p)}}^2 + \lambda_2 \|\mathbf{D}\boldsymbol{t}\|_1 + \lambda_3 \|\boldsymbol{t}\|_2^2. \tag{11}$$

Algorithm 1 Baseline Regularization

Require: $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{D}$, $\boldsymbol{l}$, $\boldsymbol{n}$, λ_1, λ_2, and λ_3.
Ensure: $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{t}}$.
1: Randomly initialize $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{t}^{(0)}$.
2: $p \leftarrow 0$.
3: **while** true **do** ▷ Outer loop: quadratic approximation
4: Compute $\mathbf{W}^{(p)}$ and $\boldsymbol{z}^{(p)}$ via (8).
5: $\tilde{\boldsymbol{t}} \leftarrow \boldsymbol{t}^{(p)}$.
6: **while** true **do** ▷ Inner loop: blockwise minimization
7: Solve for $\tilde{\boldsymbol{\beta}}$ via (10). ▷ $\boldsymbol{\beta}$-Step
8: Solve for $\tilde{\boldsymbol{t}}$ via (12). ▷ $\boldsymbol{t}$-Step
9: **if** Inner loop stopping criteria met **then**
10: $p \leftarrow p + 1$, $\boldsymbol{\beta}^{(p)} \leftarrow \tilde{\boldsymbol{\beta}}$, and $\boldsymbol{t}^{(p)} \leftarrow \tilde{\boldsymbol{t}}$.
11: **break**.
12: **end if**
13: **end while**
14: **if** Outer loop stopping criteria met **then**
15: $\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(p)}$, and $\hat{\boldsymbol{t}} \leftarrow \boldsymbol{t}^{(p)}$.
16: **return** $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{t}}$.
17: **end if**
18: **end while**

The problem in (11) is equivalent to:

$$\tilde{\boldsymbol{t}} = \arg\min_{\boldsymbol{t}} \frac{1}{2} \left\| \boldsymbol{v}^{(p)} - \boldsymbol{t} \right\|^2_{\Omega^{(p)}} + \lambda_2 \left\| \mathbf{D}\boldsymbol{t} \right\|_1, \tag{12}$$

with

$$\Omega^{(p)} = \mathbf{Z}^\top \mathbf{W}^{(p)} \mathbf{Z} + 2\lambda_3 \mathbf{I}, \quad \boldsymbol{v}^{(p)} = {\Omega^{(p)}}^{-1} \left(\mathbf{Z}^\top \mathbf{W}^{(p)} \left(\boldsymbol{z}^{(p)} - \mathbf{X}\tilde{\boldsymbol{\beta}} \right) \right).$$

The derivation from (11) to (12) is based on algebraic manipulation. Specifics are presented in the Appendix. The problem in (12) is a *blockwise weighted fused lasso signal approximator* problem. Efficient *linear time* algorithms exist for solving this type of problem [1, 2, 9, 19]. Furthermore, from (5) we notice that $\mathbf{D}$ is blockwise, so the solutions to different blocks are independent of each other. Therefore, (12) can be partitioned into various independent subproblems that can be solved *in parallel* for further speedup.

3.3 Implementation

The optimization algorithm for the BR model is summarized in Algorithm 1. Several important implementation details follow:

- To solve the problem in Step 7, we use the `glmnet` [4] package available in `R`. To solve the problem in Step 8, we use the functions from the `C` library of the `glmgen` [19] package in `R`. Both implementations are considered to be the state-of-the-art solvers for the respective subproblems.
- To avoid the divergence issue due to an ill-conditioned $\mathbf{W}^{(p)}$, we set all the diagonal elements of $\mathbf{W}^{(p)}$ that are smaller than a certain threshold, ϵ, to that threshold. In our experiments, we choose $\epsilon = 10^{-5}$. Our compact BR model by design helps to alleviate the ill-conditioned issue because a diagonal element of $\mathbf{W}^{(p)}$ represents the *cumulative* occurrence rate of adverse events during an entire interval. Ridge regularization over baseline parameters also helps to avoid small diagonal elements.

Table 1: Summary statistics of the experiment cohort

Statistics	Values
# patients	216,660
# condition (adverse event) records	1,982,000
# drug prescription records	9,089,238
Average observation duration	11.3 years

- Selection of the inner loop stopping criteria in Step 9 and the outer loop stopping criteria in Step 14 is problem-specific. We describe our choice in Section 4.4.

Our algorithmic framework shares similarities with that of `glmnet`. Both methods in the outer loop perform a quadratic approximation to a generalized linear model negative log-likelihood objective with non-smooth regularization. Both methods leverage an efficient inner loop blockwise minimization solver for the approximated problem. Therefore, both can be considered being in the family of proximal Newton methods [17, 28]. Compared with first order methods, it is well known that the proximal Newton method shares the same fast convergence rate as the usual Newton method in terms of the number of (proximal) Newton's steps needed (i.e., the number of outer loop iterations needed). However, proximal Newton methods suffer from inefficiency due to the expensive evaluations of the Hessian matrix in general. Therefore, the fact that methods under the proximal Newton framework such as `glmnet` can deliver solutions for even large-scale problems efficiently is counter-intuitive at first glance, and yet is actually attainable using an efficient inner solver [17]. Further illustrated by the experimental results to come, our algorithm provides yet another example demonstrating that the proximal Newton framework, with appropriate execution, can have the potential to handle large-scale problems effectively.

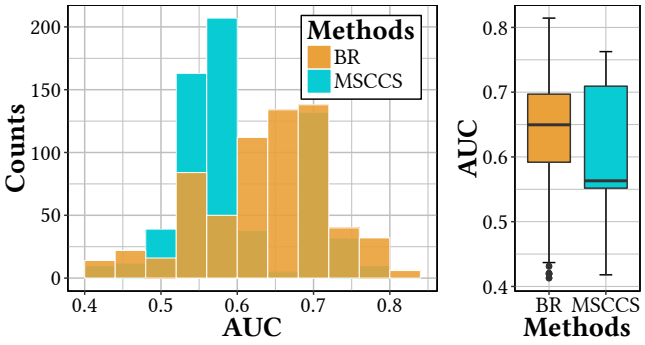
4 EXPERIMENTS

4.1 The Benchmark Task

To empirically evaluate the performance of our proposed method, we use a ground truth set of 53 drug-condition pairs generated by a selective combination of ten different drugs and nine different conditions proposed by the Observational Medical Outcomes Partnership (OMOP) [26], which was a pilot project in the U.S. aiming to conduct methodological research for the identification of ADEs from LODs. Among the 53 drug-condition pairs, 9 pairs are identified as *positive* cases (confirmed ADEs), and the remaining 44 are identified as *negative* controls. Distinguishing positive cases from the negative controls in the OMOP ground truth is widely considered to be a benchmark task for ADE discovery from LODs.

4.2 Data Source

We use the Marshfield Clinic EHR database as our data source. Being a pioneer for deploying EHR systems, Marshfield Clinic EHR database is one of the richest and the most historic in the United States, with coded diagnoses recorded as early as in 1960, and other electronic contents dating back to the 1980s [18]. We convert the diagnosis records and the drug prescription records in the EHRs to a format that is compliant with the vocabularies used in the OMOP ground truth. Following the design of MSCCS, we admit a patient



(a) Overlapping Histograms (b) Boxplot

Figure 2: Overall performance of BR and MSCCS measured by AUC among 648 different experimental configurations.

into the cohort if he or she has at least one condition of interest (adverse event) occurrence throughout the entire observation. We also further restrict our attentions to patients with at least one OMOP ground truth drug prescription record during the entire observation. Table 1 provides summary statistics of the cohort used in our experiments.

4.3 Cohort Design

We consider two important cohort design choices:

- **Risk Window Design**: a risk window is a time span that follows right after the end of a drug era during which the patient is still considered under exposure. Three types of risk windows are considered, `none`, `one month`, and `lasting`. The names of the risk windows are suggestive of their meanings.
- **Minimum Duration Design**: duration is the time length of the observation for a patient. Other than meeting the cohort admission requirement specified in Section 4.2, we admit a patient only when his or her observation duration surpasses the minimum duration threshold. We set three different minimum duration thresholds in our experiments, `none`, `three months`, and `six months`.

4.4 BR Algorithmic Design

Stopping Criteria

We denote the Euclidean norm of the difference of the two parameter vectors from the last two inner (outer) loop iterations as δ_i (δ_o). We denote the number of inner (outer) loop iterations that have run so far as c_i (c_o).

The design of the inner loop stopping criteria follow a coarse-to-fine strategy depending on how close the current outer loop iterate is to optimality. Specifically, the inner loop stopping criteria are met if any one of the following three conditions is true: (1) $\delta_o > 10$ and $\delta_i < 0.05\delta_o$; (2) $\delta_o \leq 10$ and $\delta_i < \max\left\{10^{-3}\delta_o, 10^{-4}\right\}$; (3) $c_i \geq 200$. The first criterion is useful when the current outer loop iterate is far from optimality (characterized by $\delta_o > 10$). In this case, a small number of inner loop iterations can decrease the objective effectively such that $\delta_i < 0.05\delta_o$, but further inner loop iterations do not yield much more progress. Therefore, this criterion allows the first several iterations that make significant progress, but truncates the rest that are not as effective. The second criterion determines when the inner loop stops when the current outer loop iterate is close to optimality (characterized by $\delta_o \leq 10$). In this case, the inner loop estimation needs to be more accurate to ensure that solving subsequent quadratic approximations can further decrease the objective. Therefore, the second criterion dictates that the inner loop will stop only when the estimation error is reasonably small.

The outer loop stopping criteria are met if either one of the following two conditions is true: (i) $c_o \geq 60$; (ii) $\delta_o < 10^{-4}$. Note that after each outer loop iteration, c_i is reset to 0.

Tuning Parameters

Since there are only ten different drugs available in the OMOP ground truth, the dimension of $\mathbf{X}$ is low. Therefore, we decide not to regularize $\boldsymbol{\beta}$ at all by simply setting $\lambda_1 = 0$ to decrease the complexity of the design choice space. Nonetheless, we still use `glmnet` to solve the resultant standard weighted least squares problem due to its matrix-vector friendly interface and high efficiency. We choose $\lambda_2 \in \{0.1, 0.5, 1, 2, 4, 8\}$, and $\lambda_3 \in \left\{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\right\}$. Note that to avoid overparameterization λ_2 cannot be too small. And finally, we also vary the two parameter tying strategies in Section 2.3.

The selection of λ_2, λ_3, and parameter tying strategies, along with the nine cohort design choices in Section 4.3, result in 648 different experimental configurations. Since there are nine different types of conditions, the number of BR models that are evaluated in our experiments is $648 \times 9 = 5832$.

4.5 MSCCS Algorithmic Design

An MSCCS model is an equivalent compact representation of a fixed effect Poisson regression model [34]. We therefore are able to use `glmnet` as a solver for MSCCS by learning the corresponding fixed effect Poisson regression model directly. MSCCS is a model that is only related to $\boldsymbol{\beta}$, upon which we impose a ridge penalty in our experiments. Since both BR and MSCCS share the same cohort design choices, to generate 648 experimental configurations for MSCCS as well, we use a list of 72 tuning parameters for the ridge penalty by ranging the `lambda` option in the `glmnet` function in R from 10^{-10} to 10 evenly in logarithmic scale. MSCCS without a ridge penalty is also considered. We also apply MSCCS on each of the nine different conditions, resulting in a total of 5832 different MSCCS models.

4.6 Metrics

For each of the 5832 models from both methods (BR and MSCCS), we rank the drugs in ascending order of the corresponding coefficients in the learned $\boldsymbol{\beta}$. For each of the two methods, among the models that have the *same* experimental configurations, we compute the area under curve (AUC) of receiver operating characteristics (ROC) using the OMOP ground truth and the *rankings* generated in the previous step. In this way, for both BR and MSCCS, we have 648 AUCs, each for one of the experimental configurations.

4.7 Results of Overall Performance

Since the deployed methods for ADE discovery from LODs usually reported their performances on all experimental configurations

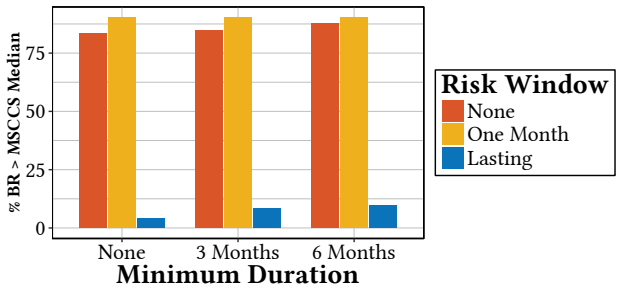
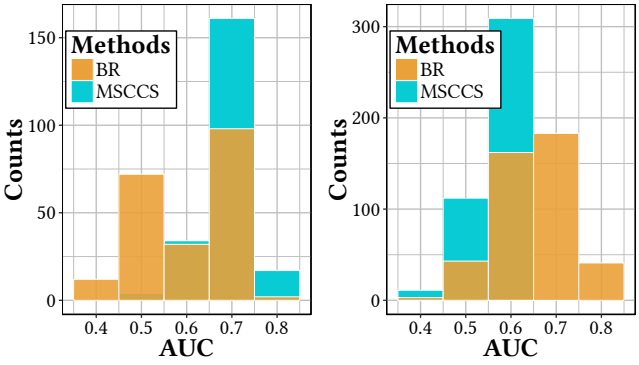


Figure 3: Proportions of BR models that outperform their median MSCCS counterparts with the same cohort design choices. Bars in different colors represent different risk window design choices. Bars in different groups represent different minimum duration design choices.



(a) Lasting Risk Window **(b) Non-Lasting Risk Window**

Figure 4: AUC Histograms of lasting and non-lasting risk window designs. Non-lasting risk windows include `none` **and** `one month`**.**

[12, 16, 21–24, 30], following this protocol, we also analyze the performances of BR and MSCCS under all of our experimental configurations.

Figure 2 visualizes the distributions of AUCs of BR and MSCCS across all 648 experimental configurations. The histogram and the box in brown represent the AUC distribution of BR and the cyan ones represent MSCCS. Compared with the AUC distribution of MSCCS, the AUC distribution of BR shifts significantly towards higher AUC intervals, with most experimental configurations achieving AUCs of more than 0.6. On the other hand, most of the experimental configurations for MSCCS achieve AUCs only between 0.5 and 0.6, which is an indication that most experimental configurations of MSCCS lack the discriminative power to separate the positive cases from the negative controls. The comparison of the overall performances between the two methods suggests that exploiting the time-varying nature of EHR data can potentially help to more accurately quantify the effects of drugs on the occurrence rate of adverse events.

4.8 Results of Various Cohort Design Choices

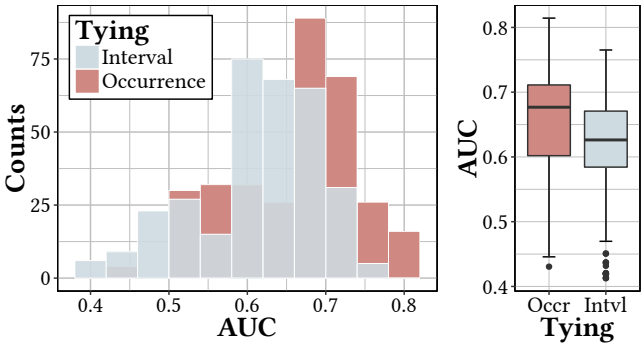
The high-variance AUC distributions of BR and MSCCS in Figure 2 motivate us to investigate under what circumstances a model will have better performance than the other. Notice that both methods share the same cohort design choices as described in Section 4.3; we therefore would like to see the effect of various cohort design choices on the performance of the two methods. To this end, for *each* of the 648 different BR models, we compute the difference of AUCs between the BR model and the MSCCS model that has the median AUC among the MSCCS models with the same cohort design choice as the BR model under consideration. We judge that the BR model outperforms the median MSCCS model with the same cohort design choice if the aforementioned difference of AUCs is larger than zero.

Risk Window

Figure 3 visualizes the proportion of BR models that outperform their median MSCCS counterparts with the same cohort design choices. The contrast visualized as bars in different colors is distinctive among the proportions of better-performing BR models with different risk window design choices: when using no risk window at all (`none`) or a one-month risk window (`one month`), the majority of BR models outperform their median MSCCS counterparts, in spite of other diversified experimental configurations that have been considered in our experiments. The proportion of better performing models under these two risk window design choices range from over 80% to an impressive 90%. As a comparison, for each cohort design choice, exactly half of the MSCCS models will outperform the corresponding median MSCCS model. Furthermore, even compared with the best performer of MSCCS models with a risk window design of `none` or `one month`, at least half of the BR models with the same risk window design choices will have better performance.

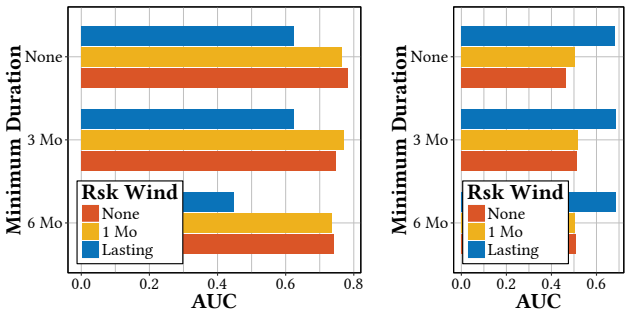
On the other hand, when the lasting risk window design choice (`lasting`) is adopted, it is more challenging for BR models to out perform their median MSCCS counterparts. A possible explanation for this phenomenon is that using a lasting risk window results in fewer and less time-varying intervals within a patient. In this setting, the data are inherently less time-varying and lack the time-dependent information that can be captured and leveraged by a BR model. Therefore, a simpler model like MSCCS might be more favorable compared to BR which might run the risk of overfitting the baseline if not regularized properly. Furthermore, recall that in our experiments we also use a ridge penalty to regularize the drug effects $\boldsymbol{\beta}$ in MSCCS models, while in BR models we do not impose any regularization over $\boldsymbol{\beta}$. The lack of time-varying intervals and the lack of regularization upon drug effects for BR models in this scenario might lead to its suboptimal performance.

Nonetheless, when being properly regularized, BR could still deliver performance comparable to MSCCS in this scenario. For example, the top performer of BR with lasting risk window gives an AUC of 0.755. The baseline is heavily regulated by $\lambda_2 = 4$ to reduce perturbational time-variability. In comparison, the best performing MSCCS model yields an AUC of 0.763.



(a) Overlapping Histograms (b) Boxplot

Figure 5: Performance of BR using the two parameter tying strategies in Section 2.3 evaluated among 648 different experimental configurations, each strategy is evaluated upon 324 configurations.



(a) BR (b) MSCCS

Figure 6: Performance of Leave-One-Condition-Out-Cross-validation (LOCOCV) among the nine cohort design choices.

Figure 4 further illustrates the performance distinction relating to the use of different risk window designs. MSCCS benefits significantly from a lasting risk window design (Figure 4a), which is consistent with the top performers reported in the literature [30]. However, assuming every drug having a lasting risk window might lead to potential model misspecification because ADEs can be caused by either long-term or short-term drug exposures [25], and many ADEs in the OMOP ground truth set are in fact acute. On the other hand, using a non-lasting (`none` or `one month`) risk window design might be more appropriate for the ground truth set in question. As shown in Figure 4b, when using BR, the ADE signals can still be effectively detected under non-lasting risk window designs.

Minimum Duration

In Figure 3, the proportions of better-performing BR models using various minimum duration design choices are represented in different groups. Given a fixed risk window, the proportions generated by the three different minimum duration thresholds do not vary as significantly as using different risk window choices.

4.9 Results of Parameter Tying

Figure 5 illustrates the effects of the two parameter tying strategies presented in Section 2.3 on the performance of various BR models. The distribution generated by occurrence tying lies in a range with higher AUCs compared with the distribution generated by interval tying. This phenomenon might be related to the clinical belief that baseline recurrence rates of adverse events tend to be different from the first occurrence rate. While occurrence tying offers a principal way to quantify this type of prior belief, interval tying might introduce redundant flexibility that focuses on perturbational baseline difference between every adjacent pair of intervals, resulting in the potential tendency to overfit the data.

4.10 Model Selection and Generalization

To demonstrate how well BR can predict unseen adverse events, for a given cohort design choice, we perform Leave-One-Condition-Out-Cross-validation (LOCOCV): for each of the nine conditions, we jointly and adaptively pick λ_2, λ_3, and the tying strategy that perform the best on the other eight conditions. In this way, we are able to use the top performer on the known ground truth to predict the unknown. We find LOCOCV to be a reasonable model selection strategy because, in essence, BR transforms the unsupervised learning of ADEs into a supervised learning problem. During learning, *none* of the ground truth label information is used. In this scenario, using LOCOCV helps us to maximize the number of training instances that can be used without worrying about the overfitting issues introduced by the ADE label information.

The AUCs of the nine different cohort design choices generated by LOCOCV are given in Figure 6. Other than under the `lasting` risk window, the AUCs of LOCOCV under other configurations exceed 0.7. In comparison, the best LOCOCV AUC from MSCCS is less than 0.7, which occurs when using a `lasting` risk window. Other configurations of MSCCS provide AUCs of around 0.5.

The reasons why we are committed to various cohort design choices are that both BR and MSCCS share the same set of cohort design choices, and that given a cohort design, the data (i.e., $\mathbf{X}$, $\boldsymbol{l}$, and $\boldsymbol{n}$) used by the two methods are exactly the same, and hence a fair comparison between the two methods can be achieved. Furthermore, in a practical setting, committing to a particular design choice can also help to facilitate the comparison of performances among multiple data sources [27].

4.11 Best Performers

In the literature of ADE discovery from LODs, it is customary to report the best performer of a method learned from a data source [12, 16, 22–24, 30]. Therefore, we also report our top performers of BR and MSCCS in our experiments: the best BR model reaches an AUC of of 0.814, with a `none` risk window, a `six months` minimum duration threshold, using occurrence tying, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$. Note that some of these configurations are somewhat different from the best configurations reported in Section 4.8, which are determined based on how well a BR model *outperforms* its best SCCS counterpart with the same cohort design choices rather than based on the absolute AUC value. The best performer of MSCCS reaches an AUC of 0.763, with a `lasting` risk window, a `three months` minimum duration threshold, and `lambda≈2.5e-3`.

5 DISCUSSION

We have proposed baseline regularization for ADE discovery from LODs. We provide an effective algorithm from the proximal Newton framework for solving the BR model and compare the performance of BR with the state-of-the-art method MSCCS in a set of diverse experimental configurations. Future research directions include running BR on other LODs for reproducibility, and accelerating the algorithm by incorporating stochasticity [15, 33, 35] and parallelism [33]. Furthermore, the current experimental configuration of BR does not consider imposing regularization upon the drug effects $\boldsymbol{\beta}$. Based on the performance gain introduced by regularizing the drug effects in MSCCS models, we speculate that introducing regularization over the drug effects in BR models will further improve its performance.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the NIH BD2K Initiative grant U54 AI117924, the NIGMS grant 2RO1 GM097618, and the P2020—Norte2020 grant NanoSTIMA/NORTE-01-0145-FEDER-00001. The authors would like to thank Dr. David Madigan from Columbia University for his constructive comment and advice. The authors would like to thank the anonymous reviewers for their helpful reviews. Sinong Geng and Zhanrong Du from the University of Wisconsin-Madison and Dr. Shijia Wang from Harvard Medical School participated in helpful discussion. They are also gratefully acknowledged.

APPENDIX

Quadratic Approximation of (3)

Let

$$f(\boldsymbol{t},\boldsymbol{\beta}) = -\log \mathcal{L}(\boldsymbol{t},\boldsymbol{\beta}) = -\boldsymbol{n}^\top (\mathbf{Z}\boldsymbol{t} + \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{l}^\top \boldsymbol{s},$$

where $\boldsymbol{s} = \exp(\mathbf{Z}\boldsymbol{t} + \mathbf{X}\boldsymbol{\beta})$. Note that $\boldsymbol{s} > \mathbf{0}$ (each component of $\boldsymbol{s}$ is strictly larger than 0) as long as $\mathbf{Z}\boldsymbol{t} + \mathbf{X}\boldsymbol{\beta}$ is bounded. For the ease of derivation, we also assume that $\begin{bmatrix}\mathbf{Z} & \mathbf{X}\end{bmatrix}$ is a column full rank matrix. In this way, an invertible Hessian of $f(\boldsymbol{t},\boldsymbol{\beta})$ can be guaranteed. The gradient and the Hessian of $f(\boldsymbol{t},\boldsymbol{\beta})$ are:

$$\nabla f(\boldsymbol{t},\boldsymbol{\beta}) = \begin{bmatrix}\mathbf{Z}^\top \\ \mathbf{X}^\top\end{bmatrix} (\mathbf{L}\boldsymbol{s} - \boldsymbol{n}), \nabla^2 f(\boldsymbol{t},\boldsymbol{\beta}) = \begin{bmatrix}\mathbf{Z}^\top \\ \mathbf{X}^\top\end{bmatrix} \mathbf{W} \begin{bmatrix}\mathbf{Z} & \mathbf{X}\end{bmatrix}, \quad (13)$$

where $\mathbf{W} = \mathbf{L}\mathbf{S}$, and $\mathbf{S} = \operatorname{diag} \boldsymbol{s}$.

At iteration p, $\boldsymbol{t}^{(p)}$ and $\boldsymbol{\beta}^{(p)}$ are given. One can show that optimizing the quadratic approximation of $f(\boldsymbol{t},\boldsymbol{\beta})$ around $\boldsymbol{t}^{(p)}$ and $\boldsymbol{\beta}^{(p)}$ is equivalent to computing a Newton's update. Using (13) and following [14], a Newton's update for $\boldsymbol{t}^{(p+1)}$ and $\boldsymbol{\beta}^{(p+1)}$ is given as:

$$\begin{bmatrix}\boldsymbol{t}^{(p+1)} \\ \boldsymbol{\beta}^{(p+1)}\end{bmatrix} = \left(\begin{bmatrix}\mathbf{Z}^\top \\ \mathbf{X}^\top\end{bmatrix} \mathbf{W}^{(p)} \begin{bmatrix}\mathbf{Z} & \mathbf{X}\end{bmatrix}\right)^{-1} \begin{bmatrix}\mathbf{Z}^\top \\ \mathbf{X}^\top\end{bmatrix} \mathbf{W}^{(p)} \boldsymbol{z}^{(p)},$$

which is the solution to the weighted least squares problem in (7), with $\boldsymbol{z}^{(p)}$ defined in (8).

Derivation from (11) to (12)

As a preparation, we state the following two algebraic facts as lemmas.

Lemma 1. Let $\boldsymbol{y}$ be an $n \times 1$ vector, let $\mathbf{X}$ be an $n \times p$ matrix, let $\boldsymbol{\beta}$ be a $p \times 1$ vector, and let $\mathbf{W}$ be a positive diagonal matrix. Then:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{W}}^2 = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} (\mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{X}\boldsymbol{\beta}) - \left(\boldsymbol{y}^\top \mathbf{W}\right) (\mathbf{X}\boldsymbol{\beta}).$$

Proof. The equation obviously holds by expanding the left hand side of the equation and removing the quantities that are not related to $\boldsymbol{\beta}$. □

Lemma 2. Let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be two $n \times 1$ vectors, let $\mathbf{X}$ be an $n \times p$ matrix, let $\boldsymbol{\beta}$ be a $p \times 1$ vector, and let $\mathbf{W}_1$, $\mathbf{W}_2$ be two positive diagonal matrices. Then:

$$\begin{aligned}&\arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{y}_1 - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{W}_1}^2 + \frac{1}{2} \|\boldsymbol{y}_2 - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{W}_2}^2 \\ &\quad = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \left\|(\mathbf{W}_1 + \mathbf{W}_2)^{-1} (\mathbf{W}_1\boldsymbol{y}_1 + \mathbf{W}_2\boldsymbol{y}_2) - \mathbf{X}\boldsymbol{\beta}\right\|_{\mathbf{W}_1+\mathbf{W}_2}^2.\end{aligned}$$

Proof. By applying Lemma 1, the quantities on both sides of the equality can be shown to be equal to:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2} (\mathbf{X}\boldsymbol{\beta})^\top (\mathbf{W}_1 + \mathbf{W}_2) (\mathbf{X}\boldsymbol{\beta}) - \left(\boldsymbol{y}_1^\top \mathbf{W}_1 + \boldsymbol{y}_2^\top \mathbf{W}_2\right) (\mathbf{X}\boldsymbol{\beta}).$$

□

We now proceed to the derivation. For convenience, we omit all the (p) superscripts and we use $\boldsymbol{\nu} = \boldsymbol{z}^{(p)} - \mathbf{X}\tilde{\boldsymbol{\beta}}$. We first show that:

$$\begin{aligned}&\arg\min_{\boldsymbol{t}} \frac{1}{2} \|\boldsymbol{\nu} - \mathbf{Z}\boldsymbol{t}\|_{\mathbf{W}}^2 + \lambda_2 \|\mathbf{D}\boldsymbol{t}\|_1 \\ &\quad = \arg\min_{\boldsymbol{t}} \frac{1}{2} \left\|\left(\mathbf{Z}^\top\mathbf{W}\mathbf{Z}\right)^{-1} \mathbf{Z}^\top\mathbf{W}\boldsymbol{\nu} - \boldsymbol{t}\right\|_{\mathbf{Z}^\top\mathbf{W}\mathbf{Z}}^2 + \lambda_2 \|\mathbf{D}\boldsymbol{t}\|_1.\end{aligned}$$

This is true because by applying Lemma 1, the quantities on both sides of the equality are equal to:

$$\arg\min_{\boldsymbol{t}} -\boldsymbol{\nu}^\top\mathbf{W}\mathbf{Z}\boldsymbol{t} + \frac{1}{2} (\mathbf{Z}\boldsymbol{t})^\top \mathbf{W} (\mathbf{Z}\boldsymbol{t}) + \lambda_2 \|\mathbf{D}\boldsymbol{t}\|_1.$$

It remains to show that

$$\arg\min_{\boldsymbol{t}} \frac{1}{2} \left\|\left(\mathbf{Z}^\top\mathbf{W}\mathbf{Z}\right)^{-1} \mathbf{Z}^\top\mathbf{W}\boldsymbol{\nu} - \boldsymbol{t}\right\|_{\mathbf{Z}^\top\mathbf{W}\mathbf{Z}}^2 + \lambda_3 \|\boldsymbol{t}\|_2^2 = \arg\min_{\boldsymbol{t}} \frac{1}{2} \|\boldsymbol{\nu} - \boldsymbol{t}\|_{\Omega}^2,$$

which is an immediate consequence of applying Lemma 2 with the fact that $\mathbf{W}_1 = \mathbf{Z}^\top\mathbf{W}\mathbf{Z}$ and $\mathbf{W}_2 = 2\lambda_3\mathbf{I}$.

REFERENCES

[1] Laurent Condat. 2013. A Direct Algorithm for 1D Total Variation Denoising. *IEEE Signal Processing Letters* (2013).

[2] P Laurie Davies and Arne Kovac. 2001. Local Extremes, Runs, Strings and Multiresolution. *Annals of Statistics* (2001).

[3] Steven Findlay. 2015. Health policy briefs: The FDA's Sentinel Initiative. *Health Affiaris* (2015).

[4] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* (2010).

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[6] Rave Harpaz, William DuMochel, and Nigam H Shah. 2015. Big Data and Adverse Drug Reaction Detection. *Clinical Pharmacology & Therapeutics* (2015).

[7] Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. 2012. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics* (2012).

[8] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, and others. 2015. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in Health Technology and Informatics* (2015).

[9] Nicholas A Johnson. 2013. A Dynamic Programming Algorithm for the Fused Lasso and l0-Segmentation. *Journal of Computational and Graphical Statistics* (2013).

[10] Zhaobin Kuang, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, and David Page. 2016. Baseline Regularization for Computational Drug Repositioning with Longitudinal Observational Data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*.

[11] David Madigan, Nandini Raghavan, William Dumouchel, Martha Nason, Christian Posse, and Greg Ridgeway. 2002. Likelihood-Based Data Squashing: A Modeling Approach to Instance Construction. *Data Mining and Knowledge Discovery* (2002).

[12] David Madigan, Martijn J Schuemie, and Patrick B Ryan. 2013. Empirical Performance of the Case–Control Method: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* (2013).

[13] Tom M Mitchell. 1997. *Machine Learning* (1 ed.). MGH.

[14] Kevin P Murphy. 2012. *Machine Learning: a Probabilistic Perspective*. MIT Press.

[15] Yu Nesterov. 2012. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization* (2012).

[16] G Niklas Norén, Tomas Bergvall, Patrick B Ryan, Kristina Juhlin, Martijn J Schuemie, and David Madigan. 2013. Empirical Performance of the Calibrated Self-Controlled Cohort Analysis within Temporal Pattern Discovery: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* (2013).

[17] Javier Pena and Ryan Tibshirani. 2016. Lecture Notes in Machine Learning 10-725/Statistics 36-725-Convex Optimization (Fall 2016). (2016).

[18] Valerie Powell, Franklin M Din, Amit Acharya, and Miguel Humberto Torres-Urquidy. 2012. *Integration of Medical and Dental Care and Patient Data*. Springer Science & Business Media.

[19] Aaditya Ramdas and Ryan J Tibshirani. 2015. Fast and Flexible ADMM Algorithms for Trend Filtering. *Journal of Computational and Graphical Statistics* (2015).

[20] Melissa A Robb, Judith A Racoosin, Rachel E Sherman, Thomas P Gross, Robert Ball, Marsha E Reichman, Karen Midthun, and Janet Woodcock. 2012. The US Food and Drug Administration's Sentinel Initiative: Expanding the Horizons of Medical Product Safety. *Pharmacoepidemiology and Drug Safety* (2012).

[21] Patrick B Ryan, David Madigan, Paul E Stang, J Marc Overhage, Judith A Racoosin, and Abraham G Hartzema. 2012. Empirical Assessment of Methods for Risk Identification in Healthcare Data: Results from the Experiments of the Observational Medical Outcomes Partnership. *Statistics in Medicine* (2012).

[22] Patrick B Ryan, Martijn J Schuemie, Susan Gruber, Ivan Zorych, and David Madigan. 2013. Empirical Performance of a New User Cohort Method: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* (2013).

[23] Patrick B Ryan, Martijn J Schuemie, and David Madigan. 2013. Empirical Performance of a Self-Controlled Cohort Method: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* (2013).

[24] Martijn J Schuemie, David Madigan, and Patrick B Ryan. 2013. Empirical Performance of LGPS and LEOPARD: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* (2013).

[25] Martijn J Schuemie, Gianluca Trifirò, Preciosa M Coloma, Patrick B Ryan, and David Madigan. 2016. Detecting Adverse Drug Reactions Following Long-Term Exposure in Longitudinal Observational Data: The Exposure-Adjusted Self-Controlled Case Series. *Statistical Methods in Medical Research* 25, 6 (2016), 2577–2592.

[26] Shawn E Simpson. 2011. *Self-Controlled Methods for Postmarketing Drug Safety Surveillance in Large-Scale Longitudinal Data*. Dissertation. Columbia University.

[27] Shawn E Simpson, David Madigan, Ivan Zorych, Martijn J Schuemie, Patrick B Ryan, and Marc A Suchard. 2013. Multiple Self-Controlled Case Series for Large-Scale Longitudinal Observational Databases. *Biometrics* (2013).

[28] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. 2012. *Optimization for Machine Learning*. Mit Press.

[29] Marc A Suchard, Shawn E Simpson, Ivan Zorych, Patrick Ryan, and David Madigan. 2013. Massive Parallelization of Serial Inference Algorithms for a Complex Generalized Linear Model. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* (2013).

[30] Marc A Suchard, Ivan Zorych, Shawn E Simpson, Martijn J Schuemie, Patrick B Ryan, and David Madigan. 2013. Empirical Performance of the Self-Controlled Case Series Design: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* (2013).

[31] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2005).

[32] Paul Tseng. 2001. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications* (2001).

[33] Stephen J Wright. 2015. Coordinate Descent Algorithms. *Mathematical Programming* (2015).

[34] Stanley Xu, Chan Zeng, Sophia Newcomer, Jennifer Nelson, and Jason Glanz. 2012. Use of Fixed Effects Models to Analyze Self-Controlled Case Series Data in Vaccine Safety Studies. *Journal of Biometrics & Biostatistics* (2012).

[35] Tuo Zhao, Mo Yu, Yiming Wang, Raman Arora, and Han Liu. 2014. Accelerated Mini-Batch Randomized Block Coordinate Descent Method. In *Advances in Neural Information Processing Systems*.