Computational Drug Repositioning Using Continuous Self-Controlled Case Series

Zhaobin Kuang[1], James Thomson[2], Michael Caldwell[3],
Peggy Peissig[4], Ron Stewart[5], David Page[6]
University of Wisconsin[1,6], Morgridge Institute[2,5], Marshfield Clinic[3,4]
zkuang@wisc.edu[1], page@biostat.wisc.edu[6],
JThomson@morgridge.org[2], RStewart@morgridgeinstitute.org[5],
caldwell.michael@marshfieldclinic.org[3], Peissig.Peggy@mcrf.mfldclin.edu[4]

ABSTRACT

Computational Drug Repositioning (CDR) is the task of discovering potential new indications for existing drugs by mining large-scale heterogeneous drug-related data sources. Leveraging the patient-level temporal ordering information between numeric physiological measurements and various drug prescriptions provided in Electronic Health Records (EHRs), we propose a Continuous Self-controlled Case Series (CSCCS) model for CDR. As an initial evaluation, we look for drugs that can control Fasting Blood Glucose (FBG) level in our experiments. Applying CSCCS to the Marshfield Clinic EHR, well-known drugs that are indicated for controlling blood glucose level are rediscovered. Furthermore, some drugs with recent literature support for the potential effect of blood glucose level control are also identified.

CCS Concepts

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

Keywords

Longitudinal Data; Self-Controlled Case Series; Computational Drug Repositioning

1. INTRODUCTION

Drug repositioning is the task of identifying new potential indications for existing drugs. This task has been steadily rising to prominence because the traditional process of *de novo* drug discovery can be slow, expensive, and risky [1]. Moreover, with the advent of the big data era, abundant data sources that collect rich drug-related information are emerging. Mining large-scale heterogeneous drug-related data sources, Computational Drug Repositioning (CDR) has become an active research area that has the potential to deliver more effective drug repositioning. There have been several comprehensive reviews in the literature on CDR [2, 3]. Many methods leverage genotypic and transcriptomic information [4, 5], as well as drug molecular structure and drug combination information [6, 7]. A prior study that used Electronic Health Records (EHRs) to validate a potential indication of *one* existing drug has also been reported [8].

We are interested in mining EHRs in order to identify a potential indication from *multiple* existing drugs simultaneously. As an initial attempt, we examine the *numeric* values of Fasting Blood Glucose (FBG) level recorded in patients' EHRs *before* and *after* some drugs are prescribed to those patients, in the hope of identifying previously unknown potential uses of drugs to control blood glucose level.

For this purpose, we extend the Self-Controlled Case Series (SCCS) [9] model that has been widely used in the Adverse Drug Reactions (ADRs) discovery community to handle *continuous* numeric response, hence the name of our model, Continuous Self-Controlled Case Series (CSCCS).

The cornerstone of a self-controlled method is an understanding of how drug prescription history will potentially influence the FBG level every time such a measurement is taken. For example, an antibiotic drug taken ten years ago might have less, if any, influence on the FBG level than an anti-diabetic drug taken a day before that FBG level is measured. To determine how long a drug can potentially influence a patient, we furthermore propose a data-driven approach that leverages change point detection [10], resulting in estimations of different time spans of influence for different drugs.

Our contributions are three-fold:

- To the best of our knowledge, this is the first translation of SCCS methodology from ADR discovery to CDR. Our work is a pilot study evaluating the use of temporal ordering information between numeric physical measurements and drug prescriptions available in EHRs for the knowledge discovery process of CDR.
- Based on the insightful observations of [11], we derive our CSCCS model from a fixed effect model and hence extend the original SCCS model to address continuous numeric response variables.
- We introduce to the CDR and ADR discovery community a data-driven approach for adaptively determining the time spans of influence of different drugs to the patients.

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.

DOI: http://dx.doi.org/10.1145/2939672.2939715

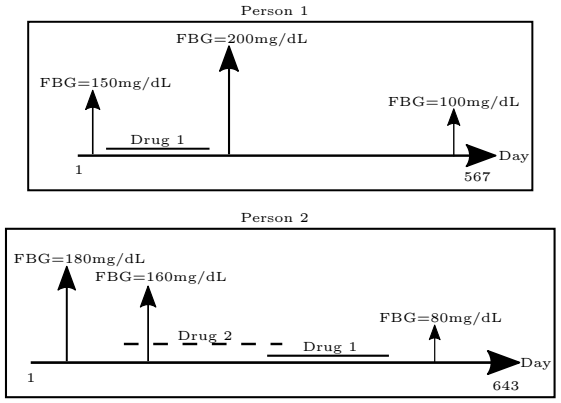


Figure 1: An example of EHRs

2. CONTINUOUS SELF-CONTROLLED CASE SERIES (CSCCS) MODEL

2.1 Notation

Figure 1 visualizes an example of health records for two patients. To confine the time span of a drug that has potential influence on that patient, we use the concept of *drug era*, which is recorded with its start date, end date and the name (or id) of the drug. We consider a patient to be under consistent influence of a drug during a drug era of that drug. However, drug era information is not readily available in most EHRs. Instead, drug prescription information with the name of a drug and the start date of the prescription is usually provided in observational data. How to construct drug eras from prescription records is a challenging and significant task for both CDR and ADR discovery [12, 13]. We provide a data-driven approach to this task in Section 4.

Measurements of FBG level might also be taken from time to time and are recorded with the date taken, as well as their numeric measurement values. We assume that at most one FBG measurement is taken for a particular patient on a particular day.

Let there be N patients with FBG measurements and M different drugs in the EHR. We construct a cohort using all the FBG measurement records as well as all the drug era records from all the N patients. Furthermore, we use a continuous random variable y_{ij}, where $i \in \{1, 2, \cdots, N\}$, $j \in \{1, 2, \cdots, J_i\}$, to denote the value of the j^{th} FBG measurement taken among a total number of J_i measurements during the observation period of the i^{th} person. Similarly, we use a binary variable x_{ijm}, $i \in \{1, 2, \cdots, N\}$, $j \in \{1, 2, \cdots, J_i\}$, $m \in \{1, 2, \cdots, M\}$ to denote the exposure status of the m^{th} drug of the i^{th} person at the date when the j^{th} FBG measurement is taken, with 1 representing exposure and 0 otherwise.

2.2 The Linear Fixed Effect Model

We treat the y_{ij}'s as the response variables and first consider the following linear regression model:

$$y_{ij}|\boldsymbol{x}_{ij} = \alpha_i + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right), \qquad (1)$$

where

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \beta_2 & \cdots & \beta_M\end{bmatrix}^\top, \quad \boldsymbol{x}_{ij} = \begin{bmatrix}x_{ij1} & x_{ij2} & \cdots & x_{ijM}\end{bmatrix}^\top,$$

α_i, which is called the *nuisance parameter*, represents the individual effect of the i^{th} person on the value of y_{ij}, invariant to day j, drug m, and other patients, and ϵ_{ij}'s are independent and identically distributed Gaussian noises with zero mean and fixed but unknown variance σ^2.

The parameter of interest in this problem is $\boldsymbol{\beta}$, which represents the effect of each of the M drugs on the response $\boldsymbol{y}$ when a patient is under the joint exposure statuses specified by $\boldsymbol{x}_{ij}$. More specifically, suppose the m^{th} component of $\boldsymbol{\beta}$, β_m, is evaluated to a negative number, that is to say, exposure to the m^{th} drug will cause the FBG level to decrease. If this drug is not known to be prescribed for lowering FBG, such a decrease is an indicator that this drug might have the potential to be repositioned to help diabetic patients control their blood glucose level, given further investigation.

In this setting, fitting a linear regression model is equivalent to solving the following least squares problem:

$$\arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{2} \left\| \boldsymbol{y} - \begin{bmatrix}\boldsymbol{Z} & \boldsymbol{X}\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \right\|_2^2, \qquad (2)$$

where

$$\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_N\end{bmatrix}^\top, \quad \boldsymbol{Z} = \operatorname{diag}\left(\mathbf{1}_1, \cdots, \mathbf{1}_N\right),$$
$$\boldsymbol{y} = \begin{bmatrix}y_{11} & \cdots & y_{1J_1} & \cdots & y_{N1} & \cdots & y_{NJ_N}\end{bmatrix}^\top,$$
$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1J_1} & \cdots & \boldsymbol{x}_{N1} & \cdots & \boldsymbol{x}_{NJ_N}\end{bmatrix}^\top,$$

where $\boldsymbol{Z}$ is a block diagonal matrix with $\mathbf{1}_i$ being a $J_i \times 1$ vector where all the components are 1. The least squares problem in (2) is a linear *fixed effect model* with $\boldsymbol{\alpha}$ being a nonrandom quantity whose i^{th} component α_i, can be interpreted as the *average* FBG measurement level of the i^{th} patient taken over time without exposing to any drugs.

2.3 Deriving the CSCCS Model from the Linear Fixed Effect Model

Like the SCCS model, the motivation behind the CSCCS model is to use only $\boldsymbol{\beta}$ as a parsimonious parameterization to predict the response vector $\boldsymbol{y}$. Inspired by the work in [11], where the equivalence between the Poisson fixed effect model and the SCCS model is established, we are able to derive the CSCCS model from the linear fixed effect model in (2) in a similar fashion. Let

$$\ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{1}{2} \left\| \boldsymbol{y} - \begin{bmatrix}\boldsymbol{Z} & \boldsymbol{X}\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \right\|_2^2.$$

We consider,

$$\frac{\partial \ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\alpha}} = \mathbf{0} \Rightarrow \boldsymbol{\alpha} = \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) = \bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{\beta}, \qquad (3)$$

where $\bar{\boldsymbol{y}}$ is an $N \times 1$ vector with the i^{th} component, $\bar{y}_i = \frac{1}{J_i}\sum_{j=1}^{J_i} y_{ij}$, and $\bar{\boldsymbol{X}}$ is an $N \times M$ matrix with the i^{th} row, $\bar{\boldsymbol{X}}_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} \boldsymbol{x}_{ij}^\top$. Substitute (3) into (2) results in the CSCCS model:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{Z}\bar{\boldsymbol{y}} - \left(\boldsymbol{X} - \boldsymbol{Z}\bar{\boldsymbol{X}}\right)\boldsymbol{\beta}\|_2^2. \qquad (4)$$

The model in (4) is in the desired form of parsimonious parameterization in that the optimization problem is defined only in the space of $\boldsymbol{\beta}$, and the nuisance parameter $\boldsymbol{\alpha}$ is eliminated.

The CSCCS model is a linear model and hence CSCCS is able to predict *continuous* response $\boldsymbol{y}$. The model is *self-controlled* in that each FBG measurement and their corresponding drug exposure statuses are adjusted by their mean

within each individual. The model also utilizes *case series* in that only cases (patients that have at least one FBG measurement) are admitted in the cohort.

CSCCS is derived from its linear fixed effect model counterpart. This derivation shares the same spirit with the equivalence between the original SCCS and the Poisson fixed effect model; in this sense, CSCCS extends SCCS to address numeric response in the new setting.

Although both models in (2) and (4) can be considered as linear models, from the perspective of implementation efficiency, the explicit form of CSCCS in (4) is of vital importance for the task of CDR using large-scale EHRs. This is because the parameter of interest in our task is $\boldsymbol{\beta}$ and the nuisance parameters do not provide direct information in evaluating the impact of a drug in changing FBG level. In the setting of large-scale EHRs, where tens of thousands of patient records might be admitted into the cohort as cases, the dimension of the nuisance parameter can potentially be very high. In this scenario, without the access to a special purpose solver for the fixed effect model, solving a model in the form of (2) using only a general purpose linear model solver can be time consuming or even infeasible. On the contrary, using the explicit form of CSCCS in (4), a general purpose linear model solver only needs to find solutions in the space of $\boldsymbol{\beta}$, a parameter whose dimension is only as large as the number of drugs available in the cohort, which is a much smaller number than the dimension of nuisance parameters.

3. CHALLENGES IN EHR DATA

Several challenges arise when we apply CSCCS to EHR data. In this section, we present the further refinements we perform on the CSCCS model presented in (4) in order to address these challenges.

3.1 High Dimensionality

EHR data is a type of high-dimensional longitudinal data. While tens of thousands of patient records might be admitted into the cohort, effects of thousands of drugs on the FBG level need to be evaluated simultaneously, introducing a high-dimensional problem. This motivates us to incorporate sparsity into our model using the penalty [14],

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{Z}\bar{\boldsymbol{y}} - \left(\boldsymbol{X} - \boldsymbol{Z}\bar{\boldsymbol{X}}\right)\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \quad (5)$$

where $\lambda > 0$ is a tuning parameter determining the level of sparsity.

The incorporation of this penalty essentially assumes that only a small portion of drugs are related to the change of FBG level, and the rest of them do not have significant effect on changing FBG level when patients are exposed to those drugs. With the L_1 penalization, most components of $\boldsymbol{\beta}$ will be evaluated to zero or a number that is close to zero. The result is, instead of evaluating the effect of *each* of the M drugs on FBG level, L_1 penalized CSCCS only selects a subset of drugs that, in some sense, are most correlated to the change of FBG level, and estimates their relative strength and direction of change among the drugs chosen.

3.2 Irregular Time Dependency

The linear fixed effect model assumes that all responses are independent of each other. The meaning of independence is two-fold. On one hand, responses from different patients are independent of each other. To explain differences across patients (e.g. some patients tend to have higher FBG levels than others in general), $\boldsymbol{\alpha}$ is used with each component representing the time-invariant effect of each patient on the response. On the other hand, responses observed at different time are independent of each other. To explain differences across time (e.g. FBG levels observed in early age *might be* lower than those in old age), a time-dependent variable that has the same value across all patients can be introduced. That is to say:

$$y_{ij}|\boldsymbol{x}_{ij} = \alpha_i + t_j + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + \epsilon_{ij}, \quad (6)$$

where t_j is the time-dependent nuisance parameter whose value depends only on the time when the j^{th} measurement is taken. If observations are recorded regularly across time, (6) defines a *two-way fixed effect model*, as opposed to the *one-way fixed effect model* defined in (2) [15].

In practice, a one-way model might be preferred over a two-way model if we assume that the heterogeneity across different individuals is much more significant than that across time. However, in the task of CDR from EHRs, this assumption might be too restrictive. To begin with, EHRs usually contain observational data of patients that are recorded over decades. Therefore, it is probable that the baseline FBG levels of patients change significantly over the years. This is especially true when some persistent FBG level altering events, such as the diagnosis of diabetes, occur to some patients. Furthermore, the length of observation periods varies dramatically among patients. Therefore, we do not have a fully observed and consistent dataset to model the set of time-dependent nuisance parameters. Last but not least, the incorporation of time-dependent nuisance parameters is proposed in a setting where data are collected regularly. With the irregular nature of EHR data, modeling time-dependent nuisance parameters directly with a classic two-way fixed effect model is impractical.

To address the aforementioned challenges without much loss in efficiency, we consider a reasonable assumption: given y_{ij} and $y_{ij'}$, where $j \neq j'$, but the dates of the two measurements taken are very close to each other, we assume the two corresponding time-dependent nuisance parameters are equal to each other, i.e. $t_j = t_{j'}$. More specifically,

$$\begin{gathered} y_{ij}|\boldsymbol{x}_{ij} = \alpha_i + t_j + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + \epsilon_{ij}, \\ y_{ij'}|\boldsymbol{x}_{ij'} = \alpha_i + t_{j'} + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij'} + \epsilon_{ij'}, \\ |d_{ij} - d_{ij'}| \leq \tau \Rightarrow t_j = t_{j'}, \end{gathered}$$

where d_{ij} and $d_{ij'}$ represent that the j^{th} and j'^{th} measurements of the i^{th} patient are taken at the d_{ij}^{th} day and $d_{ij'}^{th}$ day of the observation period, and τ is a predetermined threshold. Then,

$$\mathbb{E}\left[y_{ij} - y_{ij'}|\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}\right] = \boldsymbol{\beta}^\top\left(\boldsymbol{x}_{ij} - \boldsymbol{x}_{ij'}\right) \equiv \boldsymbol{\beta}^\top \boldsymbol{\delta}_{ij}, \quad (7)$$

where the nuisance parameters are eliminated. Therefore, the quantity in (7) depends only on $\boldsymbol{\beta}$ and the data.

Based on this formulation, we can reconstruct the CSCCS model to address irregular time dependency as follow: firstly, given τ, construct a cohort where only patients with at least a pair of FBG measurements taken within τ days are admitted; only adjacent pairs are used. Secondly, solve the following lasso problem:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{D}\boldsymbol{y} - \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \quad (8)$$

where $\boldsymbol{D}$, when multiplied with $\boldsymbol{y}$ or $\boldsymbol{X}$, generates the difference between the measurement of an earlier record and the corresponding measurement of its adjacent later-measured record of the same patient, with the constraint that the two records are collected within a time span of τ days.

Note that the model in (8) is not equivalent to the model in (5). However, the model in (8) can still be considered as a variant of CSCCS in that its parameterization is still restricted to $\boldsymbol{\beta}$, with the goal of predicting a continuous response, using data subtraction within the *same* patient as a self-controlled mechanism, and only admitting cases into the cohort. We call the the model in (8) as CSCCS for Adjacent response, or CSCCSA.

3.3 Confounding

Another challenge an algorithm must tackle is the confounding issue arises due to the complex nature of clinical observational data. In the setting of EHRs, one important confounding issue is called *confounding by co-medication*. Consider drug A and drug B, where only drug A can lower FBG level and drug B has no significant effect on changing blood sugar. However, drug B is usually prescribed with drug A. In this case, drug B can be a confounder if we only evaluate the marginal correlation between each drug and FBG level. Another confounding issue in this setting is *confounding by comorbidity*. Consider the FBG-lowering drug A given to a diabetic patient. Following the prescription of drug A, some other conditions could occur to this patient since diabetes can lead to various comorbidities [16]. To treat a newly introduced condition, drug B is prescribed to the patient. In this case, if we again consider only the marginal correlation between drug B and FBG level, one might draw the conclusion that drug B could lower FBG level since after the prescription of drug A, the FBG level has decreased.

In the two aforementioned confounding issues, drug B is called an *innocent bystander*. Like multiple SCCS [9], multiple CSCCS can effectively handle the innocent bystander confounding problem (a.k.a. Simpson's Paradox). This is because the confounder seems to spuriously correlated to the FBG level when we consider their marginal correlation. However, using a multiple linear model like CSCCS, the joint exposure statuses of both drug A and drug B can be considered simultaneously. Therefore, CSCCS might be able to identify that the decrease of FBG level occurs only when conditioning on the exposure of drug A and hence rule out drug B in the model.

In terms of addressing various confounding issues, CSCCS inherits most of the strengths and weaknesses from SCCS, due to the close relationship between the two models. While CSCCS might address reasonably well the innocent bystander confounding problem, it might not be well suited to handle confounding issues such as time-varying confounding [17]. In Section 5, we empirically evaluate the performance of CSCCS in the CDR task and illustrate how its performance is related to its capabilities of addressing various confounding issues.

4. BUILDING DRUG ERAS FROM DRUG PRESCRIPTION RECORDS

A prerequisite of CSCCS is the availability of drug era information of each drug prescribed to each patient. However, drug era information is usually not provided in most EHRs. Instead, drug prescription records of each patient are kept, usually with the name (or id) of the drug and the date of prescription. Constructing drug eras from drug prescription records is an important but challenging task for both CDR using CSCCS and ADR discovery.

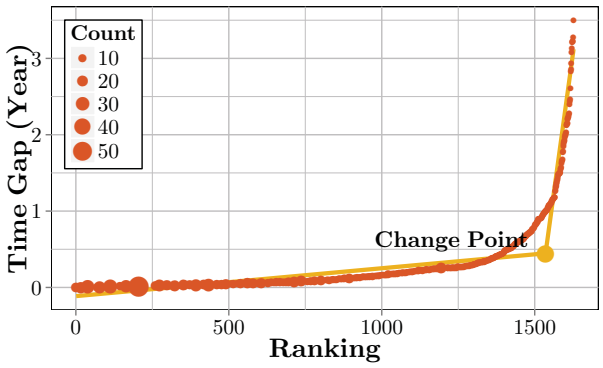


Figure 2: Time gap of Humalog in ascending order: the size of dots represents the number of time gaps that share the same value

4.1 Drug Era in Common Data Model

A heuristic proposed in the Common Data Model (CDM) [18] by Observational Medical Outcome Partnership (OMOP) is to first consider the prescription dates of each prescription record as the start date of the drug era. It then assumes that each drug era lasts n days and hence computes the end date of the drug era accordingly. Within the same patient, we assume there is only one drug prescription record of the same drug in a given date. In this way, drug eras of the same drug within each patient constructed as before start from different dates. For an adjacent pair of drug eras of the same drug within the same patient, we call the drug era that starts earlier a *former era*, and the other a *latter era*. CDM defines a parameter called *persistence window*. If the start date of the latter era, subtracted by the end date of the former era, is no larger than the persistence window, CDM merges the two drug eras into one, using the start date of the former era as the start date of the new era and the end date of the latter era as the end date of the new era. CDM tries to merge as many drug eras of the same drug within the same patient as possible in this fashion, until every resultant drug era of the same drug within the same patient is separated by more than persistence window amount of time. In CDM, both n and the persistence window are usually set to thirty days.

The intuition behind this heuristic is to build a longer drug era if the prescription date of an adjacent pair of records of the same drug are close enough to each other. A natural question to ask is *how large* the time gap between the two adjacent prescription records can be for us to still consider them close enough?

4.2 Constructing Drug Eras via Change Point Analysis

Instead of specifying a predetermined threshold on time gap as it is in CDM, we answer this question via a data-driven approach: for each drug, we compute the time gaps between all adjacent pairs of prescription records. We then sort these time gaps in ascending order. A visualization of

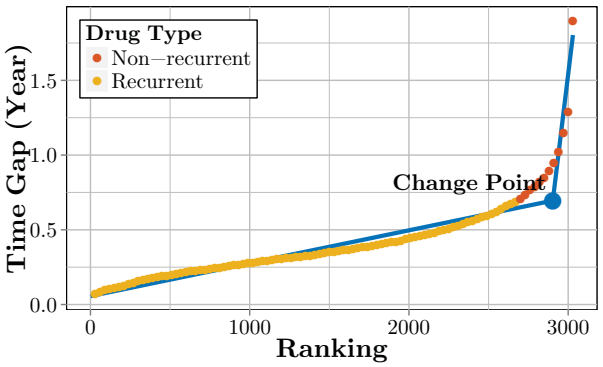


Figure 3: Change points of all drugs in the EHRs in ascending order

the values of the time gaps of Humalog against their relative rankings is given in Figure 2. From Figure 2, we notice that the distribution of time gaps can be approximated by a piecewise linear model with a change point close to the end of the sample with large time gap values. The smaller time gaps can be fitted well by the flat linear segment of the model while the larger time gaps can be fitted well by the steep linear segment. This phenomenon leads to a reasonable assumption that the smaller time gaps are sampling from a different underlying distribution than that of the larger time gaps. The smaller time gaps sampling from the same distribution correspond to the adjacent pairs of prescription records that we can consider close enough to each other to construct a lasting drug era. A threshold we can use to distinguish the two types of time gaps is the change point of the piecewise linear model.

For each drug with at least fifty prescription records in the EHRs, we perform change point detection analysis in the aforementioned fashion using `R` package `segmented`. We plot the change points of all the drugs against their relative rankings after sorting them in ascending order in Figure 3. Interestingly, there is also a change point in Figure 3. A possible explanation of the existence of a change point in Figure 3 is that in EHR data, drug prescriptions of some particular drugs are recurrent in order to battle chronic disease. For example, a diabetic patient needs long-term prescriptions of some FBG lowering drugs. On the other hand, the prescriptions of some other drugs are non-recurrent, such as antibiotics. We consider the change point in Figure 3 as a threshold to distinguish recurrent drugs from non-recurrent drugs in the EHR because a reasonable expectation is that if a drug is recurrent, the gap between an adjacent pair of prescription records of that drug from the same patient will tend not to be too large and hopefully under the change point specified in Figure 3.

We extend the heuristic provided in CDM as follow: We first denote the mean of all change point values of the recurrent drugs in the EHR as γ. For all the recurrent drugs, we set their corresponding n's and the value of their persistence windows to $\frac{\gamma}{2}$. We then set $n = 0.04$year (approximately two weeks) for all non-recurrent drugs and 0 as the value of their persistence windows.

5. EXPERIMENTS

As far as we know, our CSCCS model is the first of its kind to explicitly use temporal ordering information in EHRs for CDR. How do we evaluate the performance of a method that utilizes this type of information? As a preliminary endeavor, we try to answer this question by addressing two major challenges for our experiments.

5.1 Lack of a Baseline Method

The first challenge we need to handle is the lack of a baseline method that also utilizes temporal ordering information in an EHR for CDR. Inspired by the idea of disproportionality analysis from the pharmacovigilance literature [19], we propose the *Pairwise Mean* (PM) method as a baseline method. PM assigns a real-valued score to each of the M drugs in the EHR to represent how likely the drug decreases FBG level, and a smaller score implies a stronger decreasing tendency. The score of the m^{th} drug, s_m, is computed as follow: first, for the i^{th} patient who has FBG measurements within two years before *and* after the *first* prescription of the m^{th} drug, we compute the mean of those FBG measurements before and after the first prescription, denoted as b_{mi} and a_{mi}, respectively; second, compute s_m as:

$$s_m = \frac{1}{N_m} \sum_{i=1}^{N_m} (a_{mi} - b_{mi}) \,,$$

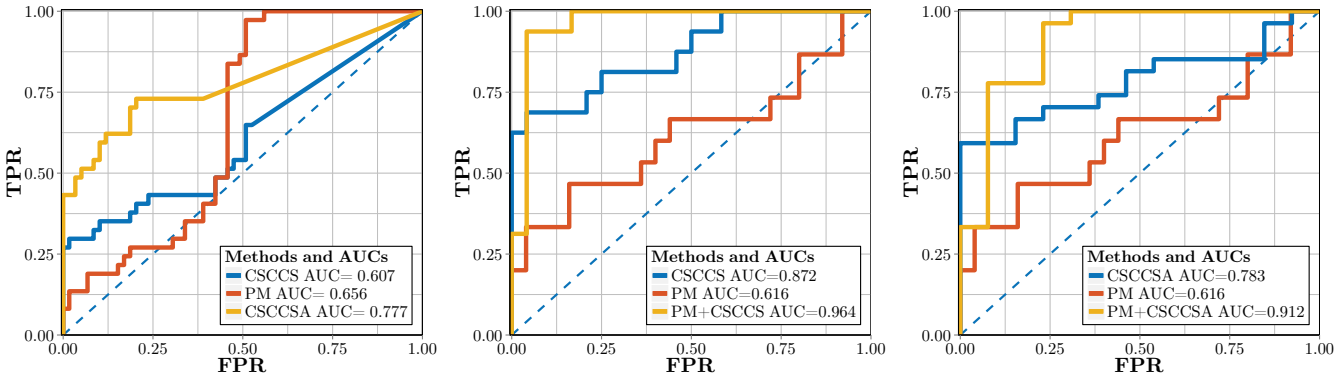
where N_m is the number of patients that have FBG measurements two years before and after the first prescription of the m^{th} drug.

5.2 Incomplete Ground Truth

Unlike the task of ADR discovery from the EHR, where numerous research efforts have been invested on developing a set of ground truth [20] drug-adverse-reaction pairs so that algorithms can be run and evaluated, we do not have access to such a ground truth set for the task of CDR from EHRs. We use Marshfield Clinic EHR as our data source and there are about two thousand drugs for evaluation. To evaluate the performance of our algorithm without knowing the glucose altering effect of every drug, we focus on the top forty most promising drugs generated by PM, CSCCS, and CSCCSA, as shown in Table 2, Table 3, and Table 4, respectively.

In these three tables, rows that are shaded in green represent the drugs commonly prescribed for lowering glucose while rows that are shaded in red represent the drugs commonly prescribed for increasing glucose. The two types of drugs in the three tables are all manually labeled. Drugs in the unshaded rows might potentially be irrelevant, or might constitute new discoveries. These drugs are discussed in further detail in Section 5.7. A summary of the number of each of the three types of drugs discovered by the three algorithms are given in Table 1.

In CSCCSA, we set τ defined in Section 3.2 to four years. In Table 2, the counts and scores are N_m's and s_m's defined in Section 5.1, while in Table 3 and Table 4, the counts are the L_1 norm of the columns in $\boldsymbol{X}$ corresponding to different drugs, and the scores are the regression coefficients of different drugs. We only consider drugs with counts greater than or equal to eight. For CSCCS and CSCCSA, we first construct drug eras using the method described in Section 4, where we determine that $\gamma = 0.34$ years. We then use a lasso penalty for variable selection to generate a long list of about two hundred drugs, and we present the top forty among those selected drugs as the short list. The number eight and forty could be tuned to optimize accuracy but



(a) PM, CSCCS, and CSCCSA (b) CSCCS (c) CSCCSA

Figure 4: ROC curves

were fixed here beforehand for practical reasons. Drugs with fewer than eight prescriptions might not have sufficient evidence to support a new use. Evaluating more than forty results per method was too large a burden for human literature review.

Table 1: A summary of three types of drugs discovered by the three algorithms

	PM	CSCCS	CSCCSA
decrease	15	16	27
increase	1	1	0
potential	24	23	13

5.3 Dataset

EHRs of 64515 patients from Marshfield Clinic are used in the CSCCS and CSCCSA experiments, providing 219306 FBG measurement records and 2980 drug candidates.

5.4 Receiver Operating Characteristic

As shown in Tables 2–4, all three methods capture a reasonable number of drugs that are prescribed for lowering glucose among their top forty candidates. We therefore consider identifying drugs prescribed for glucose-lowering as a binary classification task and use Receiver Operating Characteristics (ROC) curves as well as Area Under ROC (AUROC) to evaluate the performance of each algorithm.

We first construct the ROC curves of the three methods using the union list of drugs from Tables 2–4. The three ROC curves are presented in Figure 4(a). Since we perform variable selection in CSCCS and CSCCSA, some drugs might be assigned scores of zero and hence are considered irrelevant to the prediction of FBG level. In these cases, we put these drugs at the bottom of the union list and consider them to be identified as positive examples by the algorithms only at the very end. This results in the straight line segment of the ROC curves of CSCCS and CSCCSA at the liberal region. Figure 4(a) shows that CSCCSA has the highest AUC, outperforming CSCCS and PM by a significant margin, while PM and CSCCS have similar AUCs. However, in the more conservative region where there is drug support for all three methods, CSCCS outperforms PM while CSCCSA maintains the best performance. This phenomenon suggests that the modeling assumptions of CSCCS and CSCCSA are able to provide insights into making reasonable prediction of FBG level.

Figure 4(b) uses the forty drugs in Tables 2 and 3 to generate the ROC curves, in red for PM and in blue for CSCCS. As a comparison, we also plot the ROC curve of the following ensemble strategy: we first use the top forty drugs in Table 3 as a result of variable selection via CSCCS, then we compute the PM scores over the selected drugs. By comparing the AUCs of the three curves, we notice that the ensemble method outperforms CSCCS and PM, while CSCCS outperforms PM. Since the scores used to construct the CSCCS ROC curve are regression coefficients of drug exposure statuses under a lasso penalty, the lack of an oracle property for the lasso [21] might potentially trade off the inherent order among drugs for a sparse model. However, such a trade-off is arguably beneficial, based on the significant improvement of AUC of CSCCS compared with the AUC of PM.

Figure 4(c) is generated similarly as Figure 4(b). The ensemble of CSCCSA with PM outperforms the two individual algorithms. Although the AUC of CSCCSA is less than that of CSCCS, it is worthy to notice that all but one true positive drugs in Table 3 are discovered in Table 4 at the top fifteen positions. Other than that, CSCCSA is also able to discover twelve more true positives that CSCCS does not capture among its top forty discoveries.

5.5 Precision at K

The task of CDR from EHRs is somewhat analogous to web search. Specifically, the algorithm should select only a few drugs that have interesting unexpected effects on the response: returning too many results makes it infeasible for human experts to evaluate the potential effect of the selected drugs. This is similar to users performing web search on a search engine, where typically only the quality of the results on the first page, or the first K results, matters. Based on this observation, an algorithm with a high precision-at-K value is desirable. Figure 5 shows the precision of each of the three algorithms at different positions (K) in the task of identifying drugs prescribed for lowering glucose. CSCCSA achieves the highest performance at all positions. CSCCS outperforms PM significantly at smaller K's, but the performances of the two algorithms are similar at larger K's. This is consistent with results in Table 1, showing that CSCCSA is able to identify more prescribed drugs for lowering glucose than the other two methods. Moreover, these drugs are at the very top of Table 4. Therefore, precision-at-K provides evidence for CSCCSA's utility for CDR from EHRs.

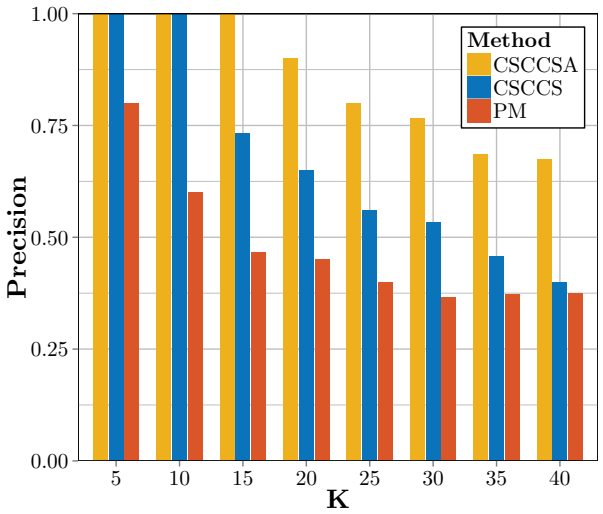


Figure 5: Precision at K of PM, CSCCS, and CSCCSA

5.6 Drugs with Known Glucose Increasing/ Decreasing Effects

From Tables 2–4, we notice that CSCCSA discovers the most number of drugs prescribed for lowering glucose among the three methods under consideration. This reaffirms our belief that CSCCSA is a promising method for CDR from EHR. Furthermore, we also notice that drugs prescribed for increasing glucose are reported in all but the table of CSCCSA.

In Table 2, sucrose is observed as a false positive using PM. Based on its count, this might be a spurious correlation in the data. This is even more probable when we consider the fact that the effect of sucrose on blood glucose level is short-term, and sucrose is not a drug that consistently enter patients' EHR for a long period of time. However, PM considers the glucose measurement records of the patients within two years before and after the first prescription of sucrose, during which many stronger confounding factors could have occurred to alter the glucose level.

In Table 3, glucagon is identified. Glucagon is given to diabetic patients that take glucose-lowering drugs to avoid hypoglycemia. However, glucagon alone is not frequently administered. Therefore, in the data, we observe the co-occurrence of glucagon with various glucose-lowering drugs. While glucagon alone increase blood glucose, combining with glucose-lowering drugs usually results in the decrease of blood sugar. On the other hand, we did not have enough data where glucagon is prescribed alone to observe the responses. Therefore, the algorithm will consider glucagon to have glucose-lowering effects since most of the time the occurrence of glucagon is accompanied by blood sugar decreasing medications. The algorithm might even consider it as a strong glucose-lowering drug because the actual glucose-lowering drugs are coded in various names in the EHR, hence dispersing the effect, while glucagon is coded only by a few different names.

5.7 Confounding and Potential Drugs

We now turn to the discussion of the drugs discovered by the three algorithms in Table 2–4 that are not prescribed for glucose increasing/decreasing. We will make use of a list providing drugs that can influence blood glucose level available in [22] to aid our evaluation process.

5.7.1 *The Blessing and the Curse of Marginal Correlations*

According to [22], in Table 2, Actigall can cause blood glucose level to increase while amphotericin B can cause blood glucose level to decrease. An interesting drug that is also brought to our attention is buderprion SR. Buderprion SR is an antidepressant prescribed for the treatment of depressive disorder. For diabetic patients with depression, buderprion SR can help to alleviate their depressive symptom, making them in a better mood. This in turn has a positive effect on better controlling blood glucose level for longer period of time [23]. PM is able to discover the blood glucose lowering effect of buderprion SR, even with a mere support of nine patients. The fact that PM considers the marginal correlation of each drug-indication pair independently makes it more likely to discover interesting drug-indication pairs with a weaker support. However, spurious correlations, especially those caused by the innocent bystander problem, are also more likely to be reported this way.

Comparing the results from Table 2 with those from Table 3 and Table 4 could justify our argument. In Table 2, Habitrol is a nicotine patch, and Monistat, Voriconazole, amphotericin B, and Hibiclens are all used to treat fungal infection. Interestingly, fungal infection is a comorbidity of diabetes [24, 25], and smokers are also more inclined to be diabetic [26]. On the other hand, we cannot find any drugs that are related to fungal infection or quitting smoke in Table 3 and Table 4. This comparison suggests that the aforementioned drugs in Table 2 generated by marginal association methods like PM might be innocent bystanders while a multiple regression approach such as CSCCS and CSCCSA might significantly help to alleviate this type of confounding issue.

5.7.2 *Potential drugs found by CSCCS and CSCCSA*

In Table 3, a study [27] indicates that enalapril helps to decrease the occurrence rate of diabetes in patients with chronic heart failure. Tricor might also have the potential to lower blood sugar level, based on the findings in [28] and [29]. Vitamin B12 is another interesting drug for consideration. In a rat model used by [30], deficiency in vitamin B12 is linked to hyperglycemia. However, blood glucose level can be decreased by providing vitamin B12. A recent study suggests that diabetic patients under metformin might experience vitamin B12 deficiency [31]. In a study on depressive patients, Zoloft, which is an antidepressant, is linked to the increase of insulin level after its prescription [32]. Zestril, which is the brand name of lisinopril, is found to inhibit high blood sugar level in rats [29]. Captopril is also reported to improve daily glucose profile among non-insulin-dependent patients [33]. However, hydralazine HCl is linked to glucose-increasing in a rat model, according to the findings in [34]. Nifedipine, verapamil HCl, and morphine sulfate can decrease blood sugar while captopril interacting with hydrochlorothiazide could cause high blood sugar, according to the list in [22]. The potential glucose-lowering drugs discovered indicate that CSCCS is a reasonable method for the task of CDR.

In Table 4, Pravachol is a member of a popular class of drugs called statins which are prescribed to lower cholesterol

Table 2: Top forty drugs: PM-Glucose

INDX	CODE	DRUG NAME	SCORE	COUNT
1	5226	LANTUS	-41.672	34
2	6646	NOVOFINE 31	-38.709	33
3	5789	METFORMIN HYDROCHLORIDE	-38.623	10
4	5806	METHENAM/MBLU/BA/SAL/ATROP/HYO	-36.710	10
5	4811	INSULIN NPH	-34.573	23
6	6652	NOVOLOG	-29.895	54
7	4336	HABITROL	-29.871	16
8	6044	MONISTAT	-29.721	14
9	9080	SURFAK	-29.655	14
10	9155	SYRNG W-NDL DISP INSUL 0.333ML	-29.439	30
11	4500	HUMULIN	-29.186	36
12	9008	SUGAR SUBSTITUTE	-28.971	10
13	10176	VORICONAZOLE	-28.538	10
14	1305	BUDEPRION SR	-27.444	9
15	8450	ROXICODONE	-27.428	12
16	9534	TRANDATE	-25.978	8
17	4802	INSULIN	-24.507	697
18	3849	FLURBIPROFEN SODIUM	-24.403	11
19	8316	REZULIN	-24.287	135
20	5257	LENALIDOMIDE	-22.875	8
21	4485	HUMALOG	-22.852	67
22	1389	CAL	-22.817	61
23	144	ACTIGALL	-22.237	36
24	8998	SUCROSE	-22.125	18
25	3843	FLUPHENAZINE HCL	-22.094	8
26	3682	FERROUS FUMARATE	-21.225	10
27	9104	SYMLINPEN 120	-20.333	12
28	1868	CHLORAMBUCIL	-20.268	14
29	4171	GLUCOTROL XL	-19.719	828
30	504	AMPHOTERICIN B	-19.672	24
31	3778	FLEXOR	-19.287	14
32	8241	REGULAR INSULIN	-19.205	39
33	824	AVANDIA	-19.140	487
34	5783	METAPROTERENOL	-18.920	10
35	4434	HIBICLENS	-18.863	10
36	5815	METH/ME BLUE/BA/PHENY/ATP/HYOS	-18.727	11
37	5010	JANUVIA	-18.716	11
38	4813	INSULIN NPL/INSULIN LISPRO	-18.515	126
39	4595	HYDROMORPHONE	-18.470	17
40	7626	POLYMYXIN B SULFATE MICRONIZED	-18.456	11

Table 3: Top forty drugs: CSCCS-Glucose

INDX	CODE	DRUG NAME	SCORE	COUNT
1	7470	PIOGLITAZONE HCL	-13.502	3075
2	8437	ROSIGLITAZONE MALEATE	-13.465	1019
3	6656	NPH HUMAN INSULIN ISOPHANE	-10.963	2874
4	4497	HUM INSULIN NPH/REG INSULIN HM	-10.869	1829
5	160	ACTOS	-7.665	1125
6	824	AVANDIA	-7.543	1239
7	4837	INSULN ASP PRT/INSULIN ASPART	-7.067	258
8	4806	INSULIN GLARGINE HUM.REC.ANLOG	-5.571	4213
9	9152	SYRING W-NDL DISP INSUL 0.5ML	-5.301	4186
10	8316	REZULIN	-3.611	444
11	3227	ENALAPRIL	-3.218	1103
12	6382	NEEDLES INSULIN DISPOSABLE	-3.148	2827
13	4970	ISOSORBIDE DINITRATE	-3.122	1220
14	9623	TRICOR	-3.119	821
15	3686	FERROUS SULFATE	-2.898	4820
16	1760	CELEXA	-2.887	1473
17	4802	INSULIN	-2.806	1526
18	4118	GLUCAGON HUMAN RECOMBINANT	-2.722	1639
19	5786	METFORMIN	-2.625	3838
20	7731	PRAVACHOL	-2.458	1700
21	2512	DARBEPOETIN ALFA IN ALBUMN SOL	-2.359	426
22	6210	MYCOPHENOLATE MOFETIL	-2.253	724
23	2830	DILTIAZEM	-2.216	1021
24	5636	MAVIK	-2.150	2242
25	4132	GLUCOPHAGE	-2.133	6736
26	4525	HYDRALAZINE HCL	-2.095	792
27	4106	GLIMEPIRIDE	-2.034	3384
28	7129	PAXIL	-2.033	2021
29	2426	CYANOCOBALAMIN (VITAMIN B-12)	-1.992	4080
30	4833	INSULIN ZINC HUMAN REC	-1.945	116
31	10392	ZOLOFT	-1.926	2417
32	6069	MORPHINE SULFATE	-1.889	899
33	10333	ZESTRIL	-1.787	2032
34	1216	BLOOD SUGAR DIAGNOSTIC	-1.665	19832
35	10199	WARFARIN SODIUM	-1.632	9223
36	3937	FOSINOPRIL SODIUM	-1.540	2660
37	6499	NIFEDIPINE	-1.524	1472
38	1003	BENAZEPRIL HCL	-1.462	1586
39	9994	VERAPAMIL HCL	-1.433	1856
40	1573	CAPTOPRIL	-1.418	1989

Table 4: Top forty drugs: CSCCSA-Glucose

INDX	CODE	DRUG NAME	SCORE	COUNT
1	4485	HUMALOG	-11.786	124
2	7470	PIOGLITAZONE HCL	-10.220	3075
3	8437	ROSIGLITAZONE MALEATE	-9.731	1019
4	4837	INSULN ASP PRT/INSULIN ASPART	-9.658	258
5	6382	NEEDLES INSULIN DISPOSABLE	-9.464	2827
6	4171	GLUCOTROL XL	-8.117	2853
7	4106	GLIMEPIRIDE	-7.940	3384
8	160	ACTOS	-7.721	1125
9	824	AVANDIA	-6.802	1239
10	9152	SYRING W-NDL DISP INSUL 0.5ML	-6.623	4186
11	4132	GLUCOPHAGE	-6.322	6736
12	4184	GLYBURIDE	-6.021	8879
13	4170	GLUCOTROL	-5.721	1259
14	4208	GLYNASE	-5.670	591
15	416	AMARYL	-5.599	2240
16	4107	GLIPIZIDE	-5.563	9993
17	844	AXID	-4.682	189
18	2830	DILTIAZEM	-4.297	1021
19	4806	INSULIN GLARGINE HUM.REC.ANLOG	-4.175	4213
20	5787	METFORMIN HCL	-4.147	19584
21	2824	DILAUDID	-4.076	39
22	5786	METFORMIN	-3.890	3838
23	7731	PRAVACHOL	-3.532	1700
24	1760	CELEXA	-3.517	1473
25	4497	HUM INSULIN NPH/REG INSULIN HM	-3.501	1829
26	9889	URSODIOL	-3.132	376
27	4813	INSULIN NPL/INSULIN LISPRO	-2.972	623
28	4133	GLUCOPHAGE XR	-2.845	765
29	6445	NEURONTIN	-2.615	1418
30	6656	NPH HUMAN INSULIN ISOPHANE	-2.500	2874
31	9379	THIAMINE HCL	-2.383	341
32	1636	CARDURA	-2.198	1079
33	1218	BLOOD SUGAR DIAGNOSTIC DRUM	-2.073	2593
34	8025	PROZAC	-2.037	1525
35	8316	REZULIN	-1.895	444
36	9136	SYRINGE & NEEDLE INSULIN 1 ML	-1.885	3542
37	4802	INSULIN	-1.812	1526
38	7674	POTASSIUM CHLORIDE	-1.779	9842
39	4804	INSULIN ASPART	-1.752	2476
40	1200	BLOOD-GLUCOSE METER	-1.719	5289

Table 5: Top forty drugs: CSCCSA-LDL

INDX	CODE	DRUG NAME	SCORE	COUNT
1	8444	ROSUVASTATIN CALCIUM	-17.052	27122
2	5368	LIPITOR	-16.908	118468
3	2395	CRESTOR	-16.234	3535
4	8720	SIMVASTATIN	-15.790	206064
5	3584	EZETIMIBE/SIMVASTATIN	-14.721	19396
6	790	ATORVASTATIN CALCIUM	-13.982	151106
7	941	BAYCOL	-12.924	1236
8	10383	ZOCOR	-11.451	26514
9	10186	VYTORIN	-9.877	9047
10	5487	LOVASTATIN	-9.238	45286
11	3583	EZETIMIBE	-8.093	32595
12	7731	PRAVACHOL	-6.729	16525
13	10336	ZETIA	-6.678	6623
14	7733	PRAVASTATIN SODIUM	-6.638	33708
15	5261	LESCOL XL	-6.358	873
16	9183	TAMOXIFEN CITRATE	-4.777	3095
17	5893	MEVACOR	-4.172	4205
18	2175	COLACE	-4.016	4349
19	9182	TAMOXIFEN	-3.764	2048
20	5260	LESCOL	-3.716	6251
21	475	AMLODIPINE/ATORVASTATIN	-2.779	1272
22	494	AMOXICILLIN/POTASSIUM CLAV	-2.495	4186
23	2110	CLOPIDOGREL BISULFATE	-2.271	50059
24	4616	HYDROXYCHLOROQUINE SULFATE	-2.240	5888
25	5281	LEVAQUIN	-2.194	1464
26	3471	ESTROGEN CON/M-PROGEST ACET	-1.929	5896
27	7496	PLAVIX	-1.471	14220
28	8225	RED YEAST RICE	-1.345	5468
29	3746	FLAGYL	-1.169	278
30	6540	NITROGLYCERIN	-1.103	94747
31	2959	DOCUSATE SODIUM	-1.084	32872
32	3475	ESTROGENS CONJUGATED	-1.033	22480
33	3686	FERROUS SULFATE	-0.990	32496
34	7768	PREMARIN	-0.969	5513
35	865	AZITHROMYCIN	-0.959	9861
36	2811	DIGOXIN	-0.908	31353
37	4132	GLUCOPHAGE	-0.779	14764
38	493	AMOXICILLIN	-0.715	11214
39	1985	CIPROFLOXACIN	-0.651	989
40	9946	VARENICLINE TARTRATE	-0.636	10794

level. Although the Food and Drug Administration (FDA) has added blood-glucose-increase warnings to all the drugs in the statin class [35], Pravachol itself has been considered to have blood-glucose lowering effects [36, 37]. The fact that CSCCSA can single out this particular drug from other statin class drug members indicates the potential of the algorithm to distinguish among similar drugs that have subtle differences. Celexa has a mild but non-significant effect on FBG level reduction in a study with seventeen depressive patients [38]. Several cases of hypoglycemia linked to the use of Neurontin have also been reported [39]. Thiamine is reported to reduce the adverse effect of hyperglycemia by inhibiting certain biological pathways [40] and deficiency of thiamine is observed in diabetic patients [41]. Cardura is found to reduce insulin resilience in a study on hypertensive patients with diabetes [42]. According to [22], Prozac

can cause both high or lower blood sugar while diltiazem is linked to low blood glucose level.

5.8 Experiments on Low-density Lipoprotein

To demonstrate the potential of our methodology, we also apply our method to predict the numeric value of low-density lipoprotein (LDL). We first construct drug eras from drug prescription records with the approach proposed in Section 4, where γ is computed as 0.36 years. We then run CSCCSA and generate a long list of about two hundred drugs. We report the top forty drugs from the list in Table 5. No confirmed false positives are discovered in the table while all the confirmed true positives are reported at the very top of the list. Some entries of hormone are discovered, which are linked to the decrease of LDL in drug/laboratory tests [43]. Interestingly, many entries of antibiotics are discovered, and all of them are classified as non-recurrent drugs by the algorithm in Section 4. This is consistent with the clinical practice that antibiotics are usually not prescribed for long-term use. Some antibiotics have also been considered to manage cholesterol level, with literature support dating back to the 1950's [44, 45, 46]. The experimental results on LDL suggest that our algorithm is not fine-tuned to boost the performance on discovering drugs that control FBG level. Instead, it is readily applicable to other important numeric clinical measurements that might lead to interesting discoveries in drug repositioning.

6. DISCUSSION

We have introduced the CSCCS model for the task of CDR using EHRs. To our best knowledge, the proposed model is the first of its kind to extensively leverage temporal ordering information from EHR to predict indications for multiple drugs at the same time. The CSCCS model extends the SCCS model that is popular in the ADR community to address a continuous response. As an initial effort, we evaluate our methodology on the task of discovering potential blood-sugar-lowering indications for a variety of drugs in a real world EHR. We develop a set of experimental evaluation methods specific to this problem in order to estimate the performance of our method. Our experimental results suggest that CSCCS can not only discover existing indications but is also able to identify potentially new use of drugs. We hence believe that CSCCS is a promising model to aid the knowledge discovery process in CDR.

Future applications and extensions of the CSCCS model are exciting. To begin with, CSCCS can be applied to a broad variety of numeric responses such as blood pressure level, cholesterol level, or body weight, to name a few. Therefore, potentially new indications of drugs to control the aforementioned important physical measurements can be examined in the same paradigm. Furthermore, many other sources of patient information, such as demographic information, diagnosis codes, other type of lab measurements, as well as interactions among all these information sources can be taken into consideration to facilitate the prediction of the physical measurement level. Last but not least, although the proposed CSCCS model in its simplest form is a linear model, the history of SCCS model development [47] could help guide on future development of CSCCS model. More complicated models can be derived from its simpler counterparts for better predictive performance in more specific and refined applications.

Acknowledgments

The authors would like to gratefully acknowledge the anonymous reviewers for their reviewing efforts and invaluable suggestions. This research is funded by the NIH BD2K Initiative grant U54 AI117924, and the NIGMS grant 2RO1 GM097618. The authors would also like to acknowledge the supports from these two grants.

References

[1] Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 2004.

[2] MR Hurle, L Yang, Q Xie, DK Rajpal, P Sanseau, and P Agarwal. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 2013.

[3] Jiao Li, Si Zheng, Bin Chen, Atul J. Butte, S. Joshua Swamidass, and Zhiyong Lu. A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics*, 2015.

[4] Justin Lamb. The connectivity map: a new tool for biomedical research. *Nature Reviews Cancer*, 2007.

[5] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 2010.

[6] Yanbin Liu, Bin Hu, Chengxin Fu, and Xin Chen. Dcdb: drug combination database. *Bioinformatics*, 2010.

[7] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, et al. Drugbank 3.0: a comprehensive resource for omics research on drugs. *Nucleic Acids Research*, 2011.

[8] Hua Xu, Melinda C Aldrich, Qingxia Chen, Hongfang Liu, Neeraja B Peterson, Qi Dai, Mia Levy, Anushi Shah, Xue Han, Xiaoyang Ruan, Min Jiang, Ying Li, Jamii St Julien, Jeremy Warner, Carol Friedman, Dan M Roden, and Joshua C Denny. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association*, 2014.

[9] Shawn E Simpson, David Madigan, Ivan Zorych, Martijn J Schuemie, Patrick B Ryan, and Marc A Suchard. Multiple self controlled case series for large scale longitudinal observational databases. *Biometrics*, 2013.

[10] Vito MR Muggeo. Estimating regression models with unknown break points. *Statistics in Medicine*, 2003.

[11] Stanley Xu, Chan Zeng, Sophia Newcomer, Jennifer Nelson, and Jason Glanz. Use of fixed effects models to analyze self-controlled case series data in vaccine safety studies. *Journal of Biometrics & Biostatistics*, 2012.

[12] Prakash M Nadkarni. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *Journal of the American Medical Informatics Association*, 2010.

[13] PB Ryan. Establishing a drug era persistence window for active surveillance. *White Papers*, 2010.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

[15] Edward W Frees. *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press, 2004.

[16] AACE. Management of common comorbidities of diabetes. http://outpatient.aace.com/type-2-diabetes/management-of-common-comorbidities-of-diabetes. (Visited on 02/06/2016).

[17] RM Daniel, SN Cousens, De BL Stavola, MG Kenward, and JAC Sterne. Methods for dealing with time dependent confounding. *Statistics in Medicine*, 2013.

[18] Stephanie J Reisinger, Patrick B Ryan, Donald J O'Hara, Gregory E Powell, Jeffery L Painter, Edward N Pattishall, and Jonathan A Morris. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of the American Medical Informatics Association*, 2010.

[19] Jean-Louis Montastruc, Agnès Sommet, Haleh Bagheri, and Maryse Lapeyre-Mestre. Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. *British Journal of Clinical Pharmacology*, 2011.

[20] OMOP. Ground truth for monitoring health outcomes of interest. http://omop.org/sites/default/files/ground%20truth.pdf, 2015. (Visited on 09/28/2015).

[21] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using l1-constrained quadratic programming. *IEEE Transactions on Information Theory*, 2009.

[22] DiabetesInControl. Drugs that can affect blood glucose levels. http://www.diabetesincontrol.com/wp-content/uploads/2010/07/www.diabetesincontrol.com_images_tools_druglistaffectingbloodglucose.pdf, 2015. (Visited on 09/28/2015).

[23] Patrick J Lustman, Monique M Williams, Gregory S Sayuk, Billy D Nix, and Ray E Clouse. Factors influencing glycemic control in type 2 diabetes during acute-and maintenance-phase treatment of major depressive disorder with bupropion. *Diabetes Care*, 2007.

[24] Jose A Vazquez and Jack D Sobel. Fungal infections in diabetes. *Infectious Disease Clinics of North America*, 1995.

[25] ADA. Skin complications. http://www.diabetes.org/living-with-diabetes/complications/skin-complications.html?referrer=https://www.google.com/. (Visited on 01/19/2016).

[26] CDC. Smoking and diabetes. http://www.cdc.gov/tobacco/campaign/tips/diseases/diabetes.html. (Visited on 01/19/2016).

[27] Emmanuelle Vermes, Anique Ducharme, Martial G Bourassa, Myriam Lessard, Michel White, and Jean-Claude Tardif. Enalapril reduces the incidence of diabetes in patients with chronic heart failure insight from the studies of left ventricular dysfunction (solvd). *Circulation*, 2003.

[28] Taner Damci, Serkan Tatliagac, Zeynep Osar, and Hasan Ilkova. Fenofibrate treatment is associated with better glycemic control and lower serum leptin and insulin levels in type 2 diabetic patients with hypertriglyceridemia. *European Journal of Internal Medicine*, 2003.

[29] Pitchai Balakumar, Rajavel Varatharajan, Ying Nyo, Raja Renushia, Devarajan Raaginey, Ann Oh, Shaikh Akhtar, Mani Rupeshkumar, Karupiah Sundram, and Sokkalingam A Dhanaraj. Fenofibrate and dipyridamole treatments in low-doses either alone or in combination blunted the development of nephropathy in diabetic rats. *Pharmacological Research*, 2014.

[30] Bacon F Chow and Howard H Stone. The relationship of vitamin b12 to carbohydrate metabolism and diabetes mellitus. *The American Journal of Clinical Nutrition*, 1957.

[31] Rose Zhao-Wei Ting, Cheuk Chun Szeto, Michael Ho-Ming Chan, Kwok Kuen Ma, and Kai Ming Chow. Risk factors of vitamin b12 deficiency in patients receiving metformin. *Archives of Internal Medicine*, 2006.

[32] Murat Kesim, Ahmet Tiryaki, Mine Kadioglu, Efnan Muci, Nuri Ihsan Kalyoncu, and Ersin Yaris. The effects of sertraline on blood lipids, glucose, insulin and hba1c levels: A prospective clinical trial on depressive patients. *Journal of Research in Medical Sciences: the Official Journal of Isfahan University of Medical Sciences*, 2011.

[33] Junichi Kodama, Shigehiro Katayama, Kiyoshi Tanaka, Akira Itabashi, Shyoji Kawazu, and Jun Ishii. Effect of captopril on glucose concentration: possible role of augmented postprandial forearm blood flow. *Diabetes Care*, 1990.

[34] Tetsuo Satoh, Shuichi Hara, Midori Takashima, and Haruo Kitagawa. Hyperglycemic effect of hydralazine in rats. *Journal of Pharmacobio-Dynamics*, 1980.

[35] FDA. Consumer updates: FDA expands advice on statin risks. http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm293330.htm, 2014. (Visited on 09/28/2015).

[36] Dilys J Freeman, John Norrie, Naveed Sattar, R Dermot G Neely, Stuart M Cobbe, Ian Ford, Christopher Isles, A Ross Lorimer, Peter W Macfarlane, James H McKillop, et al. Pravastatin and the development of diabetes mellitus evidence for a protective treatment effect in the west of scotland coronary prevention study. *Circulation*, 2001.

[37] Aleesa A Carter, Tara Gomes, Ximena Camacho, David N Juurlink, Baiju R Shah, and Muhammad M Mamdani. Risk of incident diabetes among patients treated with statins: population based study. *BMJ*, 2013.

[38] Jay D Amsterdam, Justine Shults, Nancy Rutherford, and Stanley Schwartz. Safety and efficacy of s-citalopram in patients with co-morbid major depression and diabetes mellitus. *Neuropsychobiology*, 2006.

[39] Joep HG Scholl, Rike van Eekeren, and Eugène P van Puijenbroek. Six cases of (severe) hypoglycaemia associated with gabapentin use in both diabetic and non-diabetic patients. *British Journal of Clinical Pharmacology*, 2015.

[40] Khanh vinh quoc Luong and Lan Thi Hoang Nguyen. The impact of thiamine treatment in the diabetes mellitus. *Journal of Clinical Medicine Research*, 2012.

[41] GLJ Page, David Laight, and MH Cummings. Thiamine deficiency in diabetes mellitus and the impact of thiamine replacement on glucose metabolism and vascular disease. *International Journal of Clinical Practice*, 2011.

[42] T Inukai, Y Inukai, R Matsutomo, K Okumura, K Takanashi, K Takebayashi, K Tayama, Y Aso, and Y Takemura. Clinical usefulness of doxazosin in patients with type 2 diabetes complicated by hypertension: effects on glucose and lipid metabolism. *The Journal of International Medical Research*, 2004.

[43] FDA. Premarin (conjugated estrogens tablets, usp). http://www.accessdata.fda.gov/drugsatfda_docs/label/2006/004782s147lbl.pdf. (Visited on 02/08/2016).

[44] Paul Samuel. Treatment of hypercholesterolemia with neomycin-a time for reappraisal. *New England Journal of Medicine*, 1979.

[45] YA Kesäniemi and Scott M Grundy. Turnover of low density lipoproteins during inhibition of cholesterol absorption by neomycin. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 1984.

[46] David JA Jenkins, Cyril WC Kendall, Maryam Hamidi, Edward Vidgen, Dorothea Faulkner, Tina Parker, Nalini Irani, Thomas MS Wolever, Ignatius Fong, Peter Kopplin, et al. Effect of antibiotics as cholesterol-lowering agents. *Metabolism*, 2005.

[47] David Madigan, Shawn Simpson, Wei Hua, Antonio Paredes, Bruce Fireman, and Malcolm Maclure. The self-controlled case series: Recent developments. *Submitted*, 2015. URL http://www.stat.columbia.edu/~madigan/PAPERS/DrugSafety.pdf.