PerDREP: Personalized Drug Effectiveness Prediction from Longitudinal Observational Data

Sanjoy Dey
IBM T. J. Watson Research Center
Yorktown Heights, NY
deysa@us.ibm.com

Ping Zhang
The Ohio State University
Columbus, OH
zhang.10631@osu.edu

Daby Sow
IBM T. J. Watson Research Center
Yorktown Heights, NY
sowdaby@us.ibm.com

Kenney Ng
IBM T. J. Watson Research Center
Cambridge, MA
kenney.ng@us.ibm.com

ABSTRACT

In contrast to the one-size-fits-all approach to medicine, precision medicine will allow targeted prescriptions based on the specific profile of the patient thereby avoiding adverse reactions and ineffective but expensive treatments. Longitudinal observational data such as Electronic Health Records (EHRs) have become an emerging data source for personalized medicine. In this paper, we propose a unified computational framework, called PerDREP, to predict the unique response patterns of each individual patient from EHR data. PerDREP models individual responses of each patient to the drug exposure by introducing a linear system to account for patients' heterogeneity, and incorporates a patient similarity graph as a network regularization. We formulate PerDREP as a convex optimization problem and develop an iterative gradient descent method to solve it. In the experiments, we identify the effect of drugs on Glycated hemoglobin test results. The experimental results provide evidence that the proposed method is not only more accurate than state-of-the-art methods, but is also able to automatically cluster patients into multiple coherent groups, thus paving the way for personalized medicine.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Social and professional topics** → *Medical records*; • **Computing methodologies** → *Learning linear models*.

KEYWORDS

Data Mining, Personalized Model, Healthcare Informatics

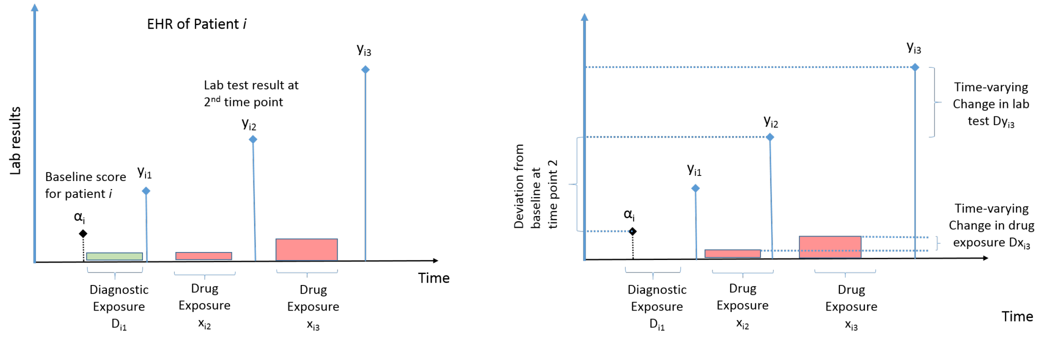
ACM Reference Format:
Sanjoy Dey, Ping Zhang, Daby Sow, and Kenney Ng. 2019. PerDREP: Personalized Drug Effectiveness Prediction from Longitudinal Observational Data. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and DataMining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3292500.3330928

1 INTRODUCTION

In contrast to one-size-fits-all medicine, personalized medicine aims to tailor treatment of a disease to the individual characteristics of each patient. This requires the ability to classify patients into subgroups with predictable response to a specific treatment. Ideally, personalized medicine will enable targeted prescription of any given treatment only to the likely responders, to avoid adverse reactions and expensive treatments to non-responders. Although there are already many examples of personalized medicine leveraging genetics/genomics information of individuals in current practice [1, 7], such information is not yet widely available in everyday clinical practice, and is insufficient since it only addresses one of many factors affecting response to medication [25]. Large-scale longitudinal observational data such as Electronic Health Records (EHRs) contain millions of patient records and thus, provides a unique opportunity to reassess the effects of a drug from many different perspectives.

Most of the existing computational methods for finding drug effectiveness from longitudinal data apply a linear fixed effect model by considering all drugs simultaneously to estimate the drug effects for a certain type of outcome of interest [21, 30, 31, 38]. To remove the effect of other clinical confounders that may vary across patients, they also leverage patient's own prior drug responses as control. Hence, these methods are called Self-Controlled Case Series (SCCS) models. Several extensions of these models (e.g., baseline regularization models presented in [20] and [14]) have been proposed to account for the variations of laboratory test results (the outcome of interest) among different patients, and to leverage the drug similarities and their therapeutic classifications for finding drug effectiveness, respectively. However, none of these studies can estimate drug effects in a personalized manner. In fact, in EHRs, there exist huge amounts of variations in terms of patients' characteristics and patients' abilities to respond to a drug. For example, one group of patients with chronic health conditions can respond to a drug differently than another group of patient with a different set of chronic health conditions [3]. Such patient heterogeneity needs to be taken into account when identifying drug effects, so that the

KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330928


EHR of Patient *i*
Lab results
Baseline score for patient *i*
α_i
y_{i1}
Lab test result at 2nd time point
y_{i2}
y_{i3}
Diagnostic Exposure D_{i1}
Drug Exposure x_{i2}
Drug Exposure x_{i3}
Time
Deviation from baseline at time point 2
α_i
y_{i1}
y_{i2}
y_{i3}
Time-varying Change in lab test Dy_{i3}
Time-varying Change in drug exposure Dx_{i3}
Diagnostic Exposure D_{i1}
Drug Exposure x_{i2}
Drug Exposure x_{i3}
Time

(a) A longitudinal EHR for a patient *i*.

(b) The effect of lab test result due to the change in drug exposure.

Figure 1: Longitudinal patient records.

obtained drugs with possible therapeutic indications and/or Adverse Drug Reactions (ADRs) can be applied in a more personalized manner during clinical decision making.

Utilizing EHR data for personalized drug response poses several challenges. First, identifying factors that can affect the patient’s ability to respond to a particular drug (i. e., heterogeneity) from longitudinal EHRs is difficult. Second, the temporal sequences when the drugs and the laboratory measurements were collected are highly irregular in nature. Third, the personalized model has to be interpretable enough, so that the local response patterns of each patient can be inferred from the model in addition to the overall global patterns. In this paper, we propose a Personalized Drug Response Prediction (PerDREP) model to identify unique response patterns of each individual patient using information from the longitudinal patient records. In particular, we use separate parameters for each individual patient for representing the drug effects on an outcome of interest. To the best of our knowledge, our model is first of its kind to account for patient heterogeneity while building interpretable predictive models for identifying drug effects from longitudinal EHR data. Our contributions in this paper can be summarized as follows:

- We introduce a linear model that can account for the patients’ heterogeneity in terms of how they respond to a particular set of drugs, which generalizes the original baseline regularization model [20].
- We incorporate several regularization schemes, so that the over-parameterization problem of the personalized drug response model is relieved substantially.
- Using one such network regularization approach using patient’s background information, we allow the clustering of patients into multiple coherent groups to learn local patterns of drug responses.
- We derive an iterative gradient descent based approach for solving the convex optimization problem.
- Finally, we demonstrate the effectiveness of our algorithm by applying the proposed method on a real-world large scale EHR data set and by conducting several case studies.

2 METHOD

2.1 Notations:

An example of a longitudinal patient record is represented in Figure 1(a). We assume that there are N patients in the EHR data with at least one record of the lab test measurement under consideration. We denote $y_{ij} \in \mathbb{R}$ as the lab test measurement of the i^{th} patient, where $i \in \{1, 2, ..., N\}$, at the j^{th} time point taken among a total number of J_i lab test measurement, i. e., $j \in \{1, 2, ..., J_i\}$. We also denote the drug exposures of M drugs for the i^{th} patient until the j^{th} time point as a vector $x_{ij} \in \mathbb{R}^M$. Each entry of this vector, x_{ijm} represents the exposure to the m^{th} drug for $m \in \{1, 2, ..., M\}$.

2.2 Problem formulations

Most of the self-controlled case series (SCCS) models assume that the measurement level of a patient obtained at a particular time is influenced by the joint effect of the exposures to drugs that the patient took until that time point, and by the baseline measurement levels due to the inherent variations among patients. Such patient specific variations of a laboratory measurement can be both time-invariant and time-dependent. For example, a time-invariant baseline effect for each individual patient can arise from existing variations among different patient groups for a particular laboratory result, which may occur due to their inherent predisposition towards certain clinical conditions (e.g., South Asian population patients have higher level of lipid profiles [2]). Furthermore, the measurements taken at one time point are not independent of the other responses that are measured at different time-points for the same patient, especially among the longitudinal observations among the large-scale temporal EHR data. Indeed, many confounding factors, both unobserved (e.g., co-morbid conditions diagnosed after the primary diagnosis for which the lab test was performed) and observed (e.g., age or weight gains) can significantly alter the laboratory responses of otherwise *healthy* subjects over such a long period of observations, irrespective of the drug exposure.

The effects of drug exposure x_{ij} towards the laboratory measurements y_{ij} along with both time-invariant and time-dependent confounding factors can be modeled using fixed effect models [31, 37] as follows:

$$y_{ij}|x_{ij} = \alpha_i + t_{ij} + \boldsymbol{\beta}^T \boldsymbol{x_{ij}} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2) \qquad (1)$$

where,

$$\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_M]^T, \quad \boldsymbol{x_{ij}} = [x_{ij1} \quad x_{ij2} \quad \cdots \quad x_{ijM}]^T,$$

Here, $\alpha_i \in \mathbb{R}$ is the patient specific, unobserved and time-invariant parameter representing the baseline effect of the i^{th} patient on the laboratory measurements y_{ij}, irrespective of time point j, drug exposures x_{ij}, and other patients. Similarly, the time-dependent parameter $t_{ij} \in \mathbb{R}$ captures the deviation of the measurement at the j^{th} point of the i^{th} patient from the baseline effect α_i. $\boldsymbol{\beta}$ is an $M \times 1$ vector with values of β_m , $m \in \{1, 2, ..., M\}$, representing the effect of m^{th} drug on the measurement of lab test. ϵ_{ij} represents independent and identically distributed Gaussian noises with zero mean and variance σ^2.

Estimating the nuisance parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha_i}$ and $\boldsymbol{t_{ij}}$ can be formulated as least square solutions of a linear fixed effect models [21, 37]:

$$\underset{\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{t}}{\arg\min} \quad \frac{1}{2} \left\| \mathrm{y} - [S \quad X \quad I] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \boldsymbol{t} \end{bmatrix} \right\|_2^2 \tag{2}$$

where,

$$\begin{aligned} \boldsymbol{\alpha} &= [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_N]^T, \\ \boldsymbol{y} &= [y_{11} \quad \cdots \quad y_{1J_1} \quad \cdots \quad y_{N1} \quad \cdots \quad y_{NJ_N}]^T, \\ \boldsymbol{X} &= [\mathbf{x}_{11} \quad \cdots \quad \mathbf{x}_{1J_1} \quad \cdots \quad \mathbf{x}_{N1} \quad \cdots \quad \mathbf{x}_{NJ_N}]^T, \\ S &= \mathrm{diag}\,(\mathbf{1}_1, \quad \mathbf{1}_2, \cdots, \quad \mathbf{1}_N), \\ \boldsymbol{t} &= \left[t_{11} \ldots t_{1J_1} \cdots t_{N1} \cdots t_{NJ_N}\right]^\top \end{aligned}$$

Here, we stack all lab test measurements of all patients into a column vector y with the dimension of J x 1, where J is the total number of lab test measurements from all patients, i.e., $J = \sum_{i=1}^{N} J_i$. Similarly, all the drug exposures are summarized in the matrix $X \in \mathbb{R}^{J \times M}$. Also, S is a block diagonal matrix with the dimension of $J \times N$, where 1_i is a $J_i \times 1$ vector with all components being 1. $\boldsymbol{\alpha}$ can represent the baseline non-random laboratory measurements of all patients. Also, $I_{J \times J}$ is the identity matrix and both $\boldsymbol{\alpha}$ and $\boldsymbol{t}$ are *nuisance parameters*, which have to be learned from the observed data.

2.3 Personalized drug response prediction

The above mentioned fixed-effect model can only estimate the baseline non-random effect of the laboratory test measurements for each person, but these methods cannot model the individual responses of each patient towards the drug exposure. The objective of our method is to find the personalized drug responses that are associated with laboratory test measurement y_{ij} that are beyond the patient specific baseline laboratory results, so that individual drug responses can be utilized for more refined decision making leading to more personalized medicine. In this paper, we extend the fixed effect models for estimating such personalized drug effect, hence the name of our model: Personalized Drug Effectiveness Prediction **(PerDREP)**. The unique assumption of this model is that there exist variations not only among the baseline measurements of laboratory results, but also among the effect of drug exposures on the laboratory test measurements for a particular patient due to *patient heterogeneity*. Therefore, the effect of drug exposures on one patient beyond the baseline effect is independent of the effect of the same drugs on other patients.

The linear fixed effect model (Eq. (1)) can be reformulated using one parameter to model the effect of one drug on one particular patient:

$$y_{ij} | x_{ij} = \alpha_i + t_{ij} + \boldsymbol{w}_i^T \boldsymbol{x}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{3}$$

where,

$$\begin{aligned} \mathbf{w}_i &= \begin{bmatrix} w_{i1} & w_{i2} & \cdots & w_{iM} \end{bmatrix}^\top, \\ \boldsymbol{W} &= \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_N \end{bmatrix}^\top \in \mathbb{R}^{N \times M} \end{aligned}$$

Here, the individual response of the i^{th} patient on the m^{th} drug is denoted by w_{im}, where $i \in [1, 2, \cdots, N]$ and $m \in [1, 2, \cdots, M]$. So, in these models, both α_i and $\boldsymbol{w_i}$ are patient-specific, but unknown time-invariant parameters representing the baseline measurement of the laboratory test and the drug effects on the same measurement respectively.

In order to solve this problem using linear least square formulation, we vectorized the drug response matrix $\mathbf{W}$ into a column vector $\boldsymbol{w} = \left[\mathbf{W}_{.1}^T \cdots \mathbf{W}_{.m}^T \cdots \mathbf{W}_{.M}^T\right]^\top$ with the dimension of $NM \times 1$, where $\mathbf{W}_{.m}$ is the m^{th} column of $\mathbf{W}$. We also rearrange the feature matrix X of Eq. (2) into a new matrix $Z = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_M \end{bmatrix}^\top$, where $Z_m \in \mathbb{R}^{J \times N}$ is a block diagonal matrix containing all the drug exposures of drug m of all patients as $Z_m = diag(\left[Z_{1m}, Z_{2m}, \cdots, Z_{Nm}\right])$.

So, if we substitute all Z_m corresponding to all drugs $m \in [1, 2, \cdots M]$, a new feature matrix Z can be obtained with the dimension of $J \times NM$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_{11} & & & & \mathbf{z}_{1M} & & \\ & \mathbf{z}_{21} & & & & \mathbf{z}_{2M} & \\ & & \ddots & \cdots & & & \ddots \\ & & & \mathbf{z}_{N1} & & & & \mathbf{z}_{NM} \end{bmatrix}$$

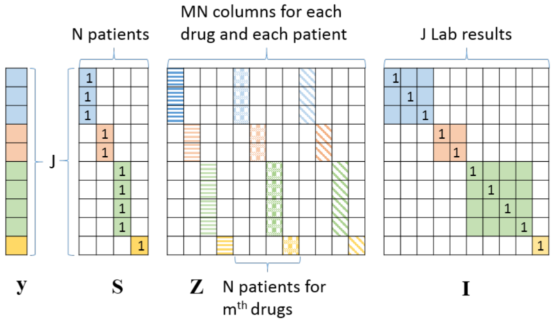

N patients
MN columns for each drug and each patient
J Lab results
J
y
S
Z
N patients for m^th drugs
I
1

Figure 2: The design matrix of Personalized Drug Response Prediction model as in Eq. (4). Each color corresponds to one patient's records.

We reformulate the personalized drug effectiveness prediction problem as a linear least square formulation as shown below:

$$\arg\min_{\boldsymbol{\alpha}, \boldsymbol{w}, \boldsymbol{t}} \quad \tfrac{1}{2} \left\| \mathrm{y} - [S \quad Z \quad I] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{w} \\ \boldsymbol{t} \end{bmatrix} \right\|_2^2 \tag{4}$$

A visual representation of the design matrix of Eq. (4) is shown in Figure 2. This least square regression problem has total number of J samples, however, the model complexity increases as we have to learn $MN + N + J$ parameters. In order to avoid over-fitting, we impose several regularizations on this model as described in next few subsections.

2.4 Time-varying Effects

We introduce a few assumptions to our PerDREP models using temporal smoothness constraints on consecutive responses of laboratory tests of each patient. Intuitively, we assume that two consecutive laboratory test results performed within a small time period do not differ significantly [20]. Let us consider two consecutive measurements y_{ij} and $y_{i(j+1)}$ of patient i taken on day π_{ij} and $\pi_{i(j+1)}$ respectively. If they are close in time, i.e., $\pi_{i(j+1)} - \pi_{ij} \leq \delta$ for a predefined threshold δ, then the changes on test measurements $y_{i(j+1)} - y_{ij}$ are either due to the drug exposures as measured by $\boldsymbol{w}_i^T x_{ij}$ or the effect of confounders within the time period δ. In both cases, we can assume $|(\alpha_i - t_{ij}) - (\alpha_i - t_{i(j+1)})| = |t_{i(j+1)} - t_{ij}|$ is small from Eq. (3). Hence, a fused lasso based regularization penalty can be incorporated [35] on the consecutive baseline parameters.

A slightly stricter assumption can be introduced into Eq. (4) by considering that consecutive test measurements that are within δ time period have the *same* baseline effect, i.e., $|\pi_{i(j+1)} - \pi_{ij}| \leq \delta \Rightarrow t_{ij} = t_{i(j+1)}$, for a small parameter δ [21]. Here, all the nuisance parameters are eliminated and the change in the laboratory test measurements only depends on $\boldsymbol{w}$. Note that although this model adopts stricter assumptions than the fused lasso based regularization, both of these models still achieve similar performances, as demonstrated in a non-personalized fixed-effect model based on Eq.(1) [20]. Since the main focus of our work is on learning personalized drug response predictions, we adopt the stricter assumptions in our model without loss of efficiency.

Given this assumption, we can reformulate our learning problem as learning the effect of changes of consecutive outputs within δ given any changes of drug exposure, as illustrated in Figure 1(b) for a sample patient record. In fact, we can construct a cohort by considering only the consecutive laboratory tests that are within δ time. Note that this cohort will also solve the issue of irregularities in the temporal dimension as described earlier. In this cohort, we can reformulate the linear learning problem as minimizing the following loss function:

$$\mathcal{L}_1 = \frac{1}{2} \left\| \boldsymbol{D}^{\delta} \boldsymbol{y} - \boldsymbol{D}^{\delta} \boldsymbol{Z} \boldsymbol{w} \right\|_2^2 \tag{5}$$

Here, $\boldsymbol{D}^{\delta}$ is a sparse matrix with dimension $s \times J$ containing only 0 or ±1 entries, where s is the total number of consecutive pairs of test measurements that are within δ. The purpose of $\boldsymbol{D}^{\delta}$ is to create a first difference matrix from the observational data, i. e., when each row of $\boldsymbol{D}^{\delta}$ is multiplied with $\boldsymbol{y}$, the new vector will contain the difference of the later measurement from the earlier measurement. For example, the difference matrix for patient i is $\boldsymbol{D}_i^{\delta} \in \mathbb{R}^{s_i \times J}$, where s_i is the total number of consecutive pairs within δ period. For each k^{th} consecutive pair $< y_{ij}, y_{i(j+1)} >$ for $k \in [1, 2, \cdots, s_i]$, the corresponding row of $\boldsymbol{D}_i^{\delta}$ will be $[0, \cdots, 0, -1, 1, 0, \cdots, 0]$ with -1 and 1 in j^{th} and $(j+1)^{th}$ positions respectively. Now, $\boldsymbol{D}^{\delta} = \text{diag}(\boldsymbol{D_1}^{\delta}, \cdots, \boldsymbol{D_N}^{\delta})$, where $s = \sum_i s_i$. For simplicity, we denote $\boldsymbol{D}^{\delta}$ as $\boldsymbol{D}$ for rest of the paper, for a given δ.

2.5 Drug Sparsity

The least square regression problem Eq. (5) has a total number of s lab results, where we have to learn a total number of MN parameters. EHR data are often high-dimensional, with large number of samples (N) considered for particular cohorts and large numbers of drugs (M) prescribed for those patients with diverse diagnostic backgrounds. However, each sample contains a small number of consecutive laboratory results that are within δ (total s results), and this can still lead to over-parameterization in Eq. (5).

To overcome this issue, we further regularize the $\boldsymbol{w}$ by imposing the regularization on the drug effectiveness within each sample so that feature selection can be performed simultaneously for improved model interpretation. The easiest way to impose sparsity is to impose an ℓ_1 penalty [34] on all drug features of all samples as shown below:

$$\mathcal{L}_2 = \lambda_1 \frac{1}{2} \sum_{i=1}^{N} \|\boldsymbol{w}_i\|_1$$

However, such heavy regularization on all parameters can lead to have many sample weights that are null due to the small number of samples available in the dataset. Instead, we would rather want to select few drugs for most of the patients. Therefore, we consider a mixed-type regularization using both ℓ_1 and ℓ_2 [19] that have been used successfully in many domains, where some predefined group structures among the variables are available. Although the definition of group is not directly applicable in our case, we can still consider each sample weight vector $\boldsymbol{w_i}$ as a group (for a total of N groups).

In our high dimensional learning case, we assume that there exists intra-group sparsity, i. e., ℓ_1 regularization is applied on individual drug exposure features within each sample (i.e., $\boldsymbol{w}_i$), while inter-group (samples) non-sparsity is achieved by imposing a ℓ_2 structure on the parameters obtained from all samples. Therefore, we want to use an $\ell_{1,2}$ or exclusive regularization [18, 41] approach to our model to impose sparsity as defined below:

$$\mathcal{L}_2 = \lambda_1 \frac{1}{2} \sum_{i=1}^{N} \|\boldsymbol{w}_i\|_1^2 \tag{6}$$

where $\lambda_1 \geq 0$ is a hyper-parameter of the model. The square of ℓ_1 in Eq. (6) will guarantee that all of the sample weights will remain non-zero (i. e., $\boldsymbol{w}_i \neq \boldsymbol{0}$).

2.6 Network Regularization

The linear least-square formulation of Eq. (5) further assumes the personalized drug responses are independent across patients. However, this assumption is not true in EHR data sets, because patients with similar backgrounds should have similar types of drug responses. For example, a particular group of patients with kidney failure may respond to a drug used to lower HbA1c in a different degree than the patient group with chronic heart diseases [24]. Based on this observation, we consider patients' demographic background and diagnosis codes as the most important determinant of the similarities among patients in terms of their background heterogeneity of having different drug responses. Consequently, we use this information to further regularize Eq. (5).

Let us consider a similarity network $\boldsymbol{R} \in \mathbb{R}^{N \times N}$, where each element $[\boldsymbol{R}]_{i,i'} = r_{ii'} \geq 0$ is a coefficient representing the relationship between each pair of patients i and i' for $i \in \{1, 2, \cdots, N\}$ and $i' \in \{1, 2, \cdots, N\}$. This graph can be computed using any similarity measure on the background information of patients i and i' such as their demographic information $G_{[i\cdot]}$ and $G_{[i'\cdot]}$, or their diagnostic profiles $P_{[i\cdot]}$ and $P_{[i'\cdot]}$, or both by combining the individual similarity scores. We assume here that $\boldsymbol{R}$ is an undirected graph (i. e., $\boldsymbol{R} = \boldsymbol{R}^T$) and the diagonal elements of $\boldsymbol{R}$ are zero, i.e., $r_{ii} = 0$ for all $i \in \{1, 2, \cdots, N\}$. Based on such relatedness of a pair of patients (i and i') $r_{ii'}$, we can impose a network regularizer on the corresponding two vectors of $\boldsymbol{w_i}$ and $\boldsymbol{w'_i}$ as follows:

$$\mathcal{L}_3 = \lambda_2 \frac{1}{2} \sum_{i,i'=1}^{N} r_{i,i'} \left\| \boldsymbol{w}_i - \boldsymbol{w}'_i \right\|_2 \tag{7}$$

where $\lambda_2 \geq 0$ is another regularization hyper-parameter.

2.7 PerDREP Model

Combining all of our assumptions described above, we arrive at the final formulation of the Personalized Drug Effectiveness Prediction model:

$$\begin{aligned} & \arg\min_{\boldsymbol{W}} \boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{L}}_1 + \boldsymbol{\mathcal{L}}_2 + \boldsymbol{\mathcal{L}}_3 \qquad (8) \\ = \quad & \arg\min_{\boldsymbol{W}} \|\boldsymbol{Dy} - \boldsymbol{DZw}\|_2^2 + \lambda_1 \sum_{i=1}^{N} \|\boldsymbol{w}_i\|_1^2 \\ & + \lambda_2 \sum_{i>i'}^{N} \sum_{i'=1}^{N-1} r_{ii'} \|\boldsymbol{w}_i - \boldsymbol{w}_{i'}\|_2 . \end{aligned}$$

Here, λ_1 and λ_2 are hyper-parameters. λ_1 controls the exclusive lasso penalty and λ_2 controls the network lasso penalty. More importantly, these two types of regularization when combined together can provide interesting model interpretations by learning multiple local predictive models [39]. If λ_2 is sufficiently large, we can efficiently cluster the samples into multiple groups based on the similarities of $\boldsymbol{w}'_i s$. More specifically, when $\|\boldsymbol{w_i} - \boldsymbol{w_{i'}}\|_2$ is quite small (preferably zero), and we can consider that the i^{th} and i'^{th} patients belong to the same cluster. At the same time outliers tend to form their own clusters that are very distant from the other normal clusters in terms of their average drug response co-efficients. Furthermore, if the λ_1 sparsity parameter is sufficiently large, then it helps selecting multiple groups of drugs where each group of drugs can correspond locally either to an individual patient or to the corresponding cluster containing the individual patient. The obtained drug responses of an individual and their localized patterns corresponding to the patient clusters can enhance the clinical interpretation of patient heterogeneity of EHR data significantly.

3 OPTIMIZATION

The PerDREP problem as formulated in Eq.(8) is a convex optimization problem where a global solution of $\boldsymbol{w}$ is available.

Lemma 1. The analytical solution the PerDREP model of Eq. (8) can be obtained by taking $\frac{\partial \mathcal{L}}{\partial w} = 0$, which can be obtained as:

$$\mathbf{w} = \left(\mathbf{Z}^\top \mathbf{D}^\top \mathbf{D}\mathbf{Z} + \mathbf{H}\right)^{-1} \mathbf{Z}^T \mathbf{D}^T \mathbf{D}\mathbf{y} \tag{9}$$

where,

$$\mathbf{H} = \lambda_1 \mathbf{F}_e + \lambda_2 \mathbf{F}_g, \quad \mathbf{F}_g = \mathbf{I}_M \otimes \mathbf{C}, \quad [\mathbf{F}_e]_{ll} = \sum_{i=1}^{N} \frac{\mathbb{I}_{il} \|\mathbf{w}_i\|_1}{[|\mathbf{w}|]_l}.$$

$$\mathbf{C} = \begin{bmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NN} \end{bmatrix},$$

$$c_{ii'} = \begin{cases} \sum_{k=1}^{n} \frac{r_{ik}}{\|\mathbf{w}_i - \mathbf{w}_k\|_2} - \frac{r_{ii'}}{\|\mathbf{w}_i - \mathbf{w}_{i'}\|_2}, & i = i', \\ -\frac{r_{ii'}}{\|\mathbf{w}_i - \mathbf{w}_{i'}\|_2}, & i \neq i'. \end{cases}$$

$$I_{i,l} = \begin{cases} 1, & \text{if } w_l \text{ belongs to patient } i \\ 0, & \text{otherwise.} \end{cases}$$

Proof.

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{w}} = \frac{\partial}{\partial w} \left(\frac{1}{2} \|\mathbf{Dy} - \mathbf{DZw}\|_2^2 \right)$$

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{w}} = -\mathbf{Z}^T \mathbf{D}^T \mathbf{D}\mathbf{y} + \mathbf{Z}^\top \mathbf{D}^\top \mathbf{D}\mathbf{Z}\mathbf{w} \tag{10}$$

The derivative of the drug sparsity regularization is [39]:

$$\frac{\partial \mathcal{L}_2}{\partial \mathbf{w}} = \frac{\partial}{\partial w} \left(\frac{1}{2} \lambda_1 \sum_{i=1}^{N} \|\boldsymbol{w}_i\|_1^2 \right) = \lambda_1 \mathbf{F}_e \mathbf{w} \tag{11}$$

Here, $\mathbf{F}_e \in \mathbb{R}^{MN \times MN}$ is a diagonal matrix. $\mathbb{I}_{il}$ is an indicator representing whether the l^{th} element in $|\mathbf{w}|$ belongs to $|\mathbf{w}_i|$.

Under $r_{ij} \geq 0$, $r_{ij} = r_{ji}, r_{ii} = 0$, the derivative of the network regularization can be obtained as [39]:

$$\frac{\partial}{\partial \mathbf{w}} \left(\sum_{i,j=1}^{N} r_{ij} \left\| \boldsymbol{w}_i - \boldsymbol{w}_j \right\|_2 \right) = 2\mathbf{C}\mathbf{W} \tag{12}$$

Now, if we perform the vectorization of the matrix W on both sides of Eq. (12), we get

$$\frac{\partial \mathcal{L}_3}{\partial \mathbf{w}} = \lambda_2 (I_M \otimes C)\mathbf{w} = \lambda_2 F_g \mathbf{w} \tag{13}$$

since, $vec(\mathbf{CWI}_M) = (\mathbf{I}_M \otimes \mathbf{C})\mathbf{w}$, where vec(*) is the vectorization operator. Here, $W = [w_1, \cdots, w_N]^T \in \mathbb{R}^{N \times M}$, $\mathbf{F}_g \in \mathbb{R}^{MN \times MN}$ is a block diagonal matrix, $\mathbf{I}_M$ is a $M \times M$ identity matrix, and $\otimes$ is the Kronecker product. Therefore, we can derive from $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0$ that

$$\mathbf{w} = (Z^T D^T DZ + H)^{-1} Z^T D^T Dy \qquad \square$$

However, in this analytical solution of *Lemma 1*, the intermediate quantities $\mathbf{F}_g$, $\mathbf{F}_e$ and $\mathbf{H}$ themselves depend on $\boldsymbol{w}$. We use a recently proposed localized lasso approach [39] to solve for the large number of parameters involved in this solution using an iterative least square method. One of the advantage of such a localized lasso optimization problem is that it does not require any tuning parameter unlike many other large-scale optimization approaches (e.g., Alternating Direction Method for Multipliers (ADMM) [4]) and guaranteed to converge to the optimal solution.

Based on the new intermediate quantity $\boldsymbol{H}$, the PerDREP optimization problem of Eq. (8) can be reformulated as optimizing the following objective function, so that $\boldsymbol{w}$ and the intermediate quantities ($\boldsymbol{F_g}$, $\boldsymbol{F_e}$, and $\boldsymbol{H}$) can be optimized iteratively.

$$\hat{\mathbb{L}} = \|\boldsymbol{Dy} - \boldsymbol{DZw}\|_2^2 + \boldsymbol{w}^\top (\lambda_1 \mathbf{F}_g^{(t)} + \lambda_2 \mathbf{F}_e^{(t)}) \boldsymbol{w} \tag{14}$$

where $\boldsymbol{F}_g^{(t)}$, $\boldsymbol{F}_e^{(t)}$ and $\boldsymbol{H}^{(t)}$ are the values of $\boldsymbol{F}_g$, $\boldsymbol{F}_e$ and $\boldsymbol{H}$ at step t. Then, $\boldsymbol{w}$ can be estimated by solving $\frac{\partial \hat{L}}{\partial w} = 0$ as below:

$$\mathbf{w}^{(t+1)} \leftarrow (\mathbf{H}^{(t)})^{-1} \mathbf{Z}^\top \mathbf{D}^\top (\mathbf{I}_n + \mathbf{DZ}(\mathbf{H}^{(t)})^{-1} \mathbf{Z}^\top \mathbf{D}^\top)^{-1} \mathbf{Dy} \tag{15}$$

Algorithm 1 Iteratively Reweighted Least Squares for localized lasso regularization

```
Data: $\mathbf{Z}$, $\mathbf{y}$, $\mathbf{D}$, $\mathbf{R}$, $\lambda_1$, $\lambda_2$
Data: $\mathbf{W}$
$t \leftarrow 0$. Initialize $\mathbf{H}^{(0)}$
repeat
    $k \longleftarrow k + 1$
    while stopping criteria not met do
        $\mathbf{w}^{(t+1)} \leftarrow (\mathbf{H}^{(t)})^{-1}\mathbf{Z}^{\top}\mathbf{D}^{\top}(\mathbf{I}_n+\mathbf{D}\mathbf{Z}(\mathbf{H}^{(t)})^{-1}\mathbf{Z}^{\top}\mathbf{D}^{\top})^{-1}\mathbf{D}\mathbf{y}$
        Compute $\mathbf{H}^{(t+1)}$ based on $\mathbf{w}^{(t+1)}$
        $t \leftarrow t + 1$
    end while
until stopping criteria not met
```

Then, $H^{(t+1)}$ is computed based on $w^{(t+1)}$ and the process can be iterated until convergence, as shown in Algorithm 1. Solving Eq. (14) leads to the global optimum solution of Eq. (8). The details are shown in the supplementary section of the paper.

4 EXPERIMENTS

We demonstrate our proposed PerDREP model using Glycated Hemoglobin (HbA1c) laboratory test measurements from our EHR data. It is widely used to measure the average blood sugar level in the body over time.

4.1 Data

We use a large-scale EHR dataset consisting of over 300,000 patients over 4 years which contains detailed time-stamped records of patient-level health events, e.g., drugs prescribed, demographics, conditions diagnosed and laboratory test results. We require that each patient have at least two HbA1c laboratory tests on different dates resulting in 14,657 patients in the HbA1c cohort. Details of the data preprocessing and cohort construction are described in the supplementary section.

4.2 Building patient similarity network

Without loss of generalizablity, we consider patients' demographics and diagnosis codes for constructing similarity network. One interesting phenomenon that we observed in the obtained network is that the similarities among the patients are not uniformly distributed, rather it has multiple modalities. Therefore, we use only the most similar patients by sparsifying the patient network using a threshold, which is denoted as *PerDREP-thre* in the rest of the paper. However, this approach often leads to only a few connected components. Alternatively, we construct a sparse fully connected network structure using a minimum spanning tree (MST) [9], which is referred as *PerDREP-MST* approach in the rest of the paper (See Supplementary section for details).

4.3 Evaluation

Lack of baseline method: To best of our knowledge, **PerDREP** is the first method of its kind to learn personalized drug response predictions for each patient by explicitly combining temporal drug exposures and patient similarity together into the same model. Since there is no available baseline for personalized models, we use one of the most efficient CSCCS method called alternative baseline regularization (ABR) [20] as the baseline model to compare against our PerDREP method. Although this baseline method cannot predict drug responses for each patient, we want to make sure that on average we do not sacrifice the global signals in term of drug responses while learning the localized and personalized patterns of the drug responses applicable only to smaller patient groups.

We perform a grid search on the exponential range of parameter values for the two hyper-parameters of our model, i.e., λ_1 and λ_2. Each combination of these hyper-parameters will return a total number of $P = | \cup_i P_i|$ drugs, where P_i represents the number of drugs for each patient. For a given P, we use the Bayesian Information criteria (BIC) [42] to assess the model fitness among those models that yield P drugs and then, select the model with minimum BIC as the best model. We also use a similar approach to optimize the ABR method, which also has two different hyper-parameters. We follow their experimental setup [21] by assigning τ=4 years.

Ground Truth: We generated two different sets of drugs which are known to treat hyperglycemia and to cause hyperglycemia side effect using MEDication Indication resource (MEDI) [36] and SIDER database (version 4.1) [22] respectively. This resulted in 77 drugs known to decrease HbA1c and 122 drugs known to increase HbA1c.

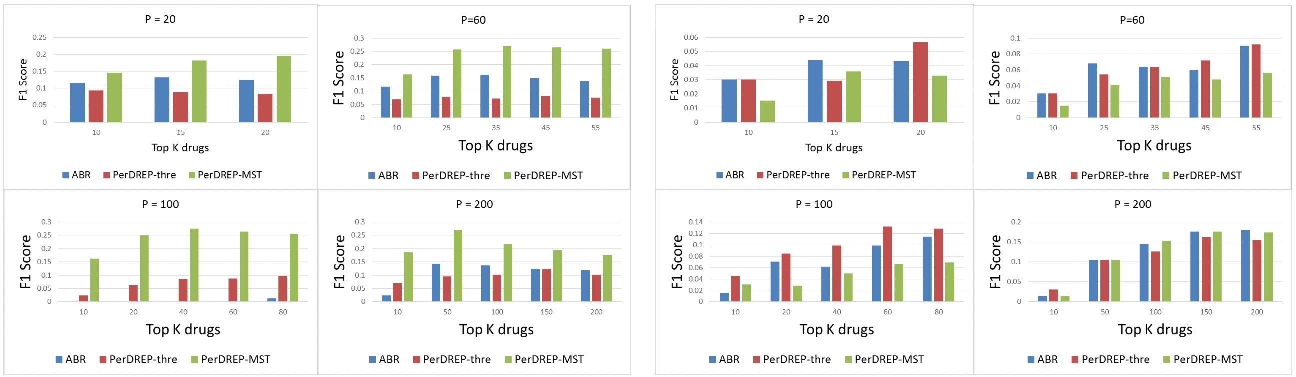
Metrics: For evaluation purposes, we use the evaluation metrics from the information retrieval domain such as precision, recall and F1-score[26] at K for two reasons. First, the ground truth data is not complete and thus the true negative drugs are not defined. Therefore, metrics such as specificity, accuracy and area under the ROC (AUC) that requires true negatives cannot be computed in this context. Second, it is only important to look at the top most drugs (K) for further evaluation by domain experts [26] and therefore, the precision and recall are computed among these K drugs. Finally, we take the harmonic average of precision and recall to get the F1-score.

5 RESULTS AND DISCUSSION

First, we compare the two versions of the PerDREP method against the baseline ABR method quantitatively. Note that the PerDREP model finds the drug effectiveness coefficients w_i for each patient separately, while ABR finds a global parameter β that represent the global signal about the drug's effectiveness. In order to generate such global coefficients for the PerDREP model for the purpose of model evaluation, we average the obtained drug coefficients W across all the patients.

To compare these methods in an unbiased manner, we consider only those models from each of the three algorithms that select similar numbers of drugs (denoted by P) while being optimized for hyper-parameters. We vary P from values $\{20, 60, 100, 200\}$ independently. For each P, we select the best models from three algorithms using the BIC criteria, and then compute the precision, recall and F1-score at K for different values of K, where $K \leq P$. Figure 3(a) shows the F1-score at K for each $P \in \{20, 60, 100, 200\}$ drugs that should decrease HbA1c values. The PerDREP-MST method significantly outperforms both the ABR and PerDREP-thre for all values of P and K.

Figure 3(b) shows the F1-score at K for P drugs that can increase HbA1c values and thus can cause potential hyperglycemia adverse drug reaction (ADR). Note that identifying drugs that cause adverse



(a) F1-score at K for P drugs that decrease HbA1c

(b) F1-score at K for P drugs that increase HbA1c

Figure 3: F1-score at K for P drugs with hyperglycemia as indication and adverse reaction.

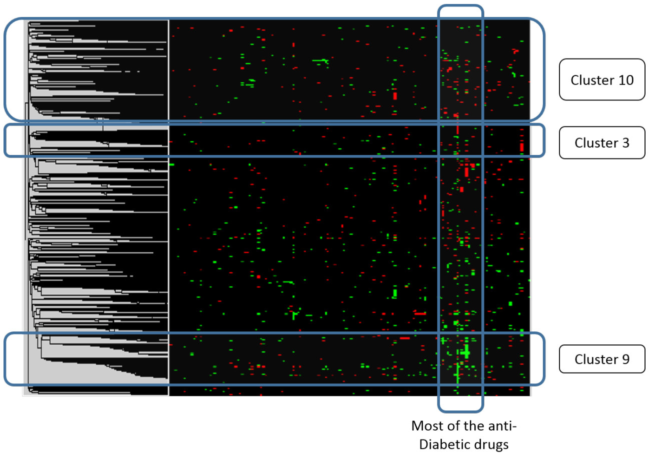


Figure 4: [Best seen in colors] Clusters obtained using a bi-clustering algorithm on both patients and drugs with K=10. The shaded box highlights three prominent clusters in both patient (row) and drug (columns) dimensions.

reactions in general is difficult and hence, the precision at K is generally lower than the precision at K for drugs that can lower HbA1c. However, the PerDREP-thre method performs as well as the ABR method, and outperforms PerDREP-MST. We note that the two PerDREP methods are complementary in nature: the MST method is better for finding effective drugs to lower HbA1c, while the threshold based technique is better at finding potential drugs that can increase HbA1c.

Identifying top-most drugs in global patterns: In order to assess the performance of our model further, we show the top 20 drugs (sorted by their scores by averaging their coefficients W) obtained from the best model selected by both PerDREP-MST and ABR model using the BIC criteria in Table 1(a). The 10 drugs with negative scores are the potential drugs to treat hyperglycemia (i.e., they lower HbA1c), while the bottom 10 drugs with positive scores can increase the HbA1c level and thus, act as potential candidates for hyperglycemia ADR. After assessing them based on the ground truth, the effective drugs (true positives) are highlighted in *pink* and the drugs that cause ADRs (true negatives) are highlighted in *green*. The *blue* ones represent false positives (i.e., with negative score) and false negatives (i.e., with positive score).

Identifying top-most drugs in local patterns: Although the PerDREP model outperforms the ABR method significantly for finding global pattern of drug effects, the real contribution of the proposed algorithm lies in discovering the localized patterns that are applicable only to a subset of patients due to patient heterogeneity. To demonstrate this capability, we cluster the trained W matrix in both dimensions, i.e, both on drugs and patients, using a bi-clustering based technique [8] with 10 clusters in each dimension. Each of the obtained bi-clusters represents a localized pattern (Figure 4). Out of these 10 clusters, there are seven prominent patient clusters (others have too few patients), each having distinct drugs as hyperglycemia indications and side-effect. After further evaluation of these clusters in terms of their distinctness, we observed interesting heterogeneity among three patient clusters (we denote them as cluster 3, 9, 10), as highlighted in Figure 4. These three clusters are very different from each other and also from the global model as shown in Table 1(b,c,d). For example, Metformin, a first-line medication for the treatment of type 2 diabetes, was selected in the top 10 effective drugs to decrease HbA1c significantly in clusters 9 and 10. However, Metformin was not selected to be effective for cluster 3. Another popular anti-diabetes drug is insulin which was selected as an effective drug globally and for clusters 3 and 10. However, interestingly, it increases HbA1c for patients in cluster 9. Similarly, Gabapentin has conflicting effects on cluster 3 and 9. These types of local patterns obtained by the PerDREP model can help generate personalized hypotheses from observational data for further clinical validation.

Case Study: Finally, we evaluate whether the effective drugs for a local cluster are associated with patients' background information, such as their disease histories and comorbidities. Specifically, we evaluate whether the effective drugs for a local cluster are associated with the specific diagnosis codes of that same cluster of patients using literature validation. As a case study, we further investigated the unique characteristics of Metformin drugs in different clusters by analyzing patients' background information. We checked the comorbidities of two different patient cohorts: patients in cluster 9 usually have diseases of the circulatory system, such as heart failure (ICD9 category of 428) and acute myocardial infarction (ICD9 category of 410). In contrast, patients in cluster 3 have more diseases of the genitourinary system, such as kidney infection (ICD9 category of 590) and chronic kidney failure (ICD9 category

(a) Global		(b) Cluster 3		(c) Cluster 9		(d) Cluster 10	
General Name	Score	General Name	Score	General Name	Score	General Name	Score
Metformin Hydrochloride	-71.8	Fluoxetine Hydrochloride	-10.7	Gabapentin	-16.4	Insulin Glargine, Recombinant	-55.5
Pioglitazone Hydrochloride	-43.5	Nystatin	-9.3	Niacin	-11.2	Glimepiride	-30.8
Glimepiride	-30.2	Promethazine Hydrochloride	-4.0	Alprazolam	-9.8	Insulin Aspart, Recombinant	-18.3
Glipizide	-28.5	Conjugated Estrogens	-2.3	Prednisone	-8.9	Metformin Hydrochloride/ rosiglitazone Maleate	-15.7
Glyburide	-24.2	Ciprofloxacin Hydrochloride	-1.9	Doxycycline Hyclate	-8.8	Azithromycin	-14.3
Sitagliptin Phosphate	-22.9	Acetaminophen/ hydrocodone Bitartrate	-1.6	Zolpidem Tartrate	-8.0	Hydrochlorothiazide/ valsartan	-10.5
Rosiglitazone Maleate	-22.7	Sildenafil Citrate	-1.0	Eszopiclone	-4.9	Sulfamethoxazole/ trimethoprim	-8.7
Insulin Glargine, Recombinant	-16.8	Citalopram Hydrobromide	-0.9	Lovastatin	-4.0	Fluticasone Propionate	-7.8
Insulin Aspart, Recombinant	-14.9	Insulin Aspart/insulin Aspart Protamine	-0.8	Glipizide/ metformin Hydrochloride	-4.0	Metformin Hydrochloride/ pioglitazone Hydrochloride	-7.1
Lisinopril	-14.6	Zolpidem Tartrate	-0.8	Polyethylene Glycol 3350	-4.0	Lactulose	-6.2
Ibuprofen	9.7	Omeprazole	20.2	Acetaminophen/ hydrocodone Bitartrate	41.6	Lisinopril	104.6
Diltiazem Hydrochloride	10.2	Tetracycline Hydrochloride	9.9	Insulin Glargine, Recombinant	37.9	Simvastatin	61.9
Amoxicillin	10.4	Finasteride	8.3	Ibuprofen	27.1	Lovastatin	52.7
Glucose Meter	11.4	Ramipril	4.4	Insulin Lispro, Recombinant	24.9	Hydrochlorothiazide	46.0
Gemfibrozil	11.8	Acetaminophen/ propoxyphene Napsylate	2.1	Furosemide	21.9	Glyburide	44.0
Ketoconazole	12.5	Acetylcysteine	0.9	Fenofibrate	18.2	Atenolol	33.2
Albuterol Sulfate	14.5	Gabapentin	0.8	Amlodipine Besylate	17.2	Cephalexin	29.6
Cephalexin	15.1	Cephalexin	0.7	Potassium Chloride	15.3	Valsartan	28.5
Prednisone	15.4	Glipizide	0.7	Olanzapine	7.6	Levothyroxine Sodium	28.4
Fluconazole	21.1	Esterified Estrogens	0.6	Carvedilol	7.0	Ibuprofen	28.0

Table 1: Top 20 ranked drugs for all patients (a) and three important patient clusters (b,c,d). The first 10 drugs with negative co-efficients decrease HbA1c, while the last 10 drugs increase HbA1c.

of 585). The observation is consistent with clinical guidelines: heart failure patients who use Metformin have better outcomes than those on other anti-diabetic agents [24]; insulin is not effective for heart failure patients in controlling their blood sugar levels, and will even increase their risk of mortality [33]. However, the use of Metformin for patients with kidney diseases is controversial [16]; instead, insulin is usually used for those comorbidities (insulin significantly decreases HbA1c for those patients in cluster 3).

6 RELATED WORK

Drug effectiveness prediction: Finding therapeutic effects of drugs has been studied using the properties of drugs such as chemical properties [11], chemical-protein interactome [13, 23], or the properties of diseases such as disease gene networks[28], disease gene expression [32]. Also, some studies [15, 40] leverage one or more such properties of drugs and diseases for predicting therapeutic indications of newly developed drugs or by re-purposing already existing drugs. A different group of studies use similar types data sources for predicting adverse drug reactions (ADRs) [6, 27]. However, none of these studies used observational data for predicting the therapeutic indication or ADRs. Recently, observational data such as EHRs, spontaneous reporting system data, health examination survey have been used by a few studies to find potential ADRs [17] and therapeutic indications [5]. However, none of these studies take the longitudinal information of patient’s observational data nor build the personalized models taking the same patient’s prior history as control.

Machine learning techniques for mining drug effects from longitudinal EHRs: Most of the existing machine learning studies aimed to apply a linear model on the longitudinal patient history for estimating drug effects for a certain type of outcome of interest such as decreased cancer risk [30], decreased fasting blood glucose [21], or increased risk of ADRs [31]. Since they leveraged the patient’s own previous drug responses as control, these methods are called Self-Controlled Case Series (SCCS) models. Recently, a baseline regularization model [20] has been proposed to utilize the drug histories over time using a baseline parameter in the model which can account for the variations of laboratory test results (the outcome of interest) among different patients. Furthermore, another recent method [14] extends the baseline regularization model by leveraging drug similarities and therapeutic classifications to guide the optimization of identifying drug effectiveness on the laboratory test results. However, none of these studies can estimate drug effects in a personalized manner. In totally different contexts of mining EHR data for predicting disease phenotypes, a few recent studies aim to predict individualized [38] and local patterns [10, 12] of treatment responses. However, these methods are not applicable for predicting unexpected drug responses and characterizing responsive patients. In addition, none of these studies address the challenges of EHR data including defining patient’s heterogeneity from EHR data, addressing the irregular temporal nature of EHR data and model interpretability.

7 CONCLUSION

We have introduced a personalized drug response prediction model to identify unique response patterns of each individual patient using longitudinal patient record data. Experimental results suggest that the proposed method is not only more accurate than state-of-the-art methods, but is also able to cluster the patients automatically into multiple groups that are clinically coherent. We believe that the method can potentially aid the knowledge discovery process for personalized medicine. Our method can further be extended to include diagnostic and genetic data explicitly into the model for learning patient heterogeneity.

ACKNOWLEDGMENTS

The authors would like to thank Zhaobin Kuang for providing the **R** code of the *ABR* method and answering questions we had during the development of the project.

REFERENCES

[1] Akram Alyass, Michelle Turcotte, and David Meyre. 2015. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics* 8, 33 (2015).

[2] Ozlem Bilen, Ayeesha Kamal, and Salim S Virani. 2016. Lipoprotein abnormalities in South Asians and its association with cardiovascular disease: Current state and future directions. *World journal of cardiology* 8, 3 (2016), 247.

[3] Cynthia Boyd, Jonathan Darer, Chad Boult, et al. 2005. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *The Journal of the American Medical Association (JAMA)* 294, 6 (2005), 716–724.

[4] Stephen Boyd, Neal Parikh, Eric Chu, et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3, 1 (2011), 1122.

[5] Adam S Brown, Danielle Rasooly, and Chirag J Patel. 2018. Leveraging Population-Based Clinical Quantitative Phenotyping for Drug Repositioning. *CPT: pharmacometrics & systems pharmacology* 7, 2 (2018), 124–129.

[6] D-S Cao, N Xiao, Y-J Li, et al. 2015. Integrating multiple evidence sources to predict adverse drug reactions based on a systems pharmacology model. *CPT: pharmacometrics & systems pharmacology* 4, 9 (2015), 498–506.

[7] Rui Chen and Michael Snyder. 2012. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 5, 1 (2012), 73–82.

[8] Yizong Cheng and George M Church. 2000. Biclustering of expression data.. In *Ismb*, Vol. 8. 93–103.

[9] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. 2009. *Introduction to algorithms*. MIT press.

[10] Sanjoy Dey, Jacob Cooner, Connie W Delaney, Joanna Fakhoury, Vipin Kumar, Gyorgy Simon, Michael Steinbach, Jeremy Weed, and Bonnie L Westra. 2015. Mining patterns associated with mobility outcomes in home healthcare. *Nursing research* 64, 4 (2015), 235–245.

[11] Sanjoy Dey, Heng Luo, Achille Fokoue, Jianying Hu, and Ping Zhang. 2018. Predicting adverse drug reactions through interpretable deep learning framework. *BMC bioinformatics* 19, 21 (2018), 476.

[12] Sanjoy Dey, Gyorgy Simon, Bonnie Westra, Michael Steinbach, and Vipin Kumar. 2014. Mining interpretable and predictive diagnosis codes from multi-source electronic health records. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, 1055–1063.

[13] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. 2018. Interpretable Drug Target Prediction Using Deep Neural Representation. In *IJCAI*. 3371–3377.

[14] Mohamed Ghalwash, Ying Li, Ping Zhang, and Jianying Hu. 2017. Exploiting electronic health records to mine drug effects on laboratory test results. In *ACM on Conference on Information and Knowledge Management*. 1837–1846.

[15] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* 7, 1 (2011), 496.

[16] Allison Hahr and Mark Molitch. 2015. Management of diabetes mellitus in patients with chronic kidney disease. *Clinical Diabetes and Endocrinology* 1, 2 (2015).

[17] Rave Harpaz, William DuMouchel, Paea LePendu, et al. 2013. Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System. *Clinical Pharmacology & Therapeutics* 93, 6 (2013), 539–546.

[18] Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding. 2014. Exclusive Feature Learning on Arbitrary Structures via ell_1, 2-norm. In *Advances in Neural Information Processing Systems*. 1655–1663.

[19] Matthieu Kowalski. 2009. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis* 27, 3 (2009), 303–324.

[20] Zhaobin Kuang, James Thomson, Michael Caldwell, et al. 2016. Baseline regularization for computational drug repositioning with longitudinal observational data. In *IJCAI: proceedings of the conference*, Vol. 2016. NIH Public Access, 2521.

[21] Zhaobin Kuang, James Thomson, Michael Caldwell, et al. 2016. Computational drug repositioning using continuous self-controlled case series. In *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining*, Vol. 2016. NIH Public Access, 491.

[22] Michael Kuhn, Ivica Letunic, Lars Jensen, and Peer Bork. 2015. The SIDER database of drugs and side effects. *Nucleic Acids Research* 44, D1 (2015), D1075–D1079.

[23] Heng Luo, Ping Zhang, Xi Hang Cao, et al. 2016. Dpdr-cpi, a server that predicts drug positioning and drug repositioning via chemical-protein interactome. *Scientific reports* 6 (2016), 35996.

[24] Michael MacDonald, Dean Eurich, Sumit Majumdar, et al. 2010. Treatment of type 2 diabetes and outcomes in patients with heart failure: a nested case control study from the UK general practice research database. *Diabetes Care* 33, 6 (2010), 1213–1218.

[25] Urs Meyer. 2004. Pharmacogenetics - five decades of therapeutic lessons from genetic diversity. *Nature Reviews Genetics* 5, 9 (2004), 669–676.

[26] Emir Muñoz, Vít Nováček, and Pierre-Yves Vandenbussche. 2016. Using drug similarities for discovery of possible adverse reactions. In *AMIA Annual Symposium Proceedings*, Vol. 2016. AMIA, 924.

[27] Emir Muñoz, Vít Nováček, and Pierre-Yves Vandenbussche. 2017. Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. *Briefings in bioinformatics* (2017).

[28] Sunghong Park, Dong-gi Lee, and Hyunjung Shin. 2017. Network mirroring for drug repositioning. *BMC medical informatics and decision making* 17, 1 (2017), 55.

[29] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. 2008. The matrix cookbook. *Technical University of Denmark* 7, 15 (2008), 510.

[30] Rikje Ruiter, Loes E Visser, Myrthe PP van Herk-Sukel, et al. 2012. Lower risk of cancer in patients on metformin in comparison with those on sulfonylurea derivatives. *Diabetes care* 35, 1 (2012), 119–124.

[31] 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.

[32] Marina Sirota, Joel T Dudley, Jeewon Kim, Annie P Chiang, et al. 2011. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine* 3, 96 (2011), 96ra77–96ra77.

[33] Stephanie Smooke, Tamara Horwich, and Gregg Fonarow. 2005. Insulin-treated diabetes is associated with a marked increase in mortality in patients with advanced heart failure. *American Heart Journal* 149, 1 (2005), 168–174.

[34] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[35] 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)* 67, 1 (2005), 91–108.

[36] Wei-Qi Wei, Robert Cronin, Hua Xu, et al. 2013. Development and evaluation of an ensemble resource linking medications to their indications. *J Am Med Inform Assoc* 20, 5 (2013), 954–961.

[37] 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), 006.

[38] Yanbo Xu, Yanxun Xu, and Suchi Saria. 2016. A Bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference*. 282–300.

[39] Makoto Yamada, Takeuchi Koh, Tomoharu Iwata, John Shawe-Taylor, and Samuel Kaski. 2017. Localized Lasso for High-Dimensional Regression. In *Artificial Intelligence and Statistics*. 325–333.

[40] Ping Zhang, Fei Wang, and Jianying Hu. 2014. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In *AMIA Annual Symposium Proceedings*, Vol. 2014. American Medical Informatics Association, 1258.

[41] Yang Zhou, Rong Jin, and Steven Hoi. 2010. Exclusive lasso for multi-task feature selection. In *International conference on artificial intelligence and statistics*. 988–995.

[42] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. 2007. On the "degrees of freedom" of the lasso. *The Annals of Statistics* 35, 5 (2007), 2173–2192.

8 SUPPLEMENTARY SECTION

8.1 Additional Proof of Optimization algorithm

Theorem 1: Algorithm 1 converges to the optimal solution Eq. (8).

Proof. The solution of the alternative loss function $\hat{\mathcal{L}}$ of Eq. (14) can be obtained by solving $\frac{\partial \hat{L}}{\partial w} = 0$ as below [29]:

$$\mathbf{w}^{(t+1)} \leftarrow (\mathbf{H}^{(t)})^{-1}\mathbf{Z}^{\top}\mathbf{D}^{\top}(\mathbf{I}_n + \mathbf{D}\mathbf{Z}(\mathbf{H}^{(t)})^{-1}\mathbf{Z}^{\top}\mathbf{D}^{\top})^{-1}\mathbf{D}\mathbf{y}$$

Using the updated rules of Eq. (15) to solve the alternate formulation $\hat{\mathcal{L}}$, it can be shown that [18, 39] the new objective function of Eq. (14) is monotonically decreasing in each iteration:

$$\mathcal{L}(w^{(t+1)}) - \mathcal{L}(w^{(t)}) \leq \hat{\mathcal{L}}(w^{(t+1)}) - \hat{\mathcal{L}}(w^{(t)}) \leq 0.$$

Based on this relationship, the algorithm will also monotonically decrease the original formulation of Eq. (8). At convergence, H^* will satisfy Eq. (9), which implies that the corresponding solution, w^* will be the global solution of the convex optimization problem Eq. (8). Therefore, Algorithm 1 will converge to the global optimal solution of Eq. (8). □

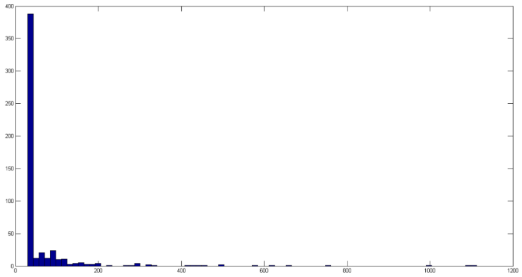


Figure 5: The frequency of the drug era rank of a particular drug.

8.2 Data Preparation

The pharmacy data in the EHR database contain information such as the national drug code (NDC) of medication, the date of medication supplied and the number of days supplied. We map NDCs to their generic names using a resource named Redbook[1], which is a propriety resource providing the mapping between NDCs and their generic names. The diagnosis information in the EHR database contain information such as International Classification of Disease (ICD)-9 codes[2] and their diagnosis dates. In the study, we used the first three digits of the ICD9 codes (i.e., ICD9 category) to represent patents' comorbidities, which yields 1026 unique ICD9 categories in total.

8.2.1 *Cohort Construction.*

A key challenge in longitudinal EMRs is how to find the "drug exposure era" from the raw prescription records. In this study, we assume a drug can have the effectiveness period of n days for each patient, which will lead to multiple drug eras starting with a unique date for a particular drug and a particular patient. Next, we merge all the consecutive drug eras that are too close to each other, i.e, the ending date of the first era is within a persistent window of the start date of the second era of the same drug. These two parameters, namely n and *persistent window*, were learned from the observational data using statistics of the length of the drug era (Figure 5). Intuitively, we wanted to detect the optimal point when the lengths of the drug era change drastically. For simplicity, we assume that these lengths of drug era are bi-modal, i.e., the optimal point for each drug can be obtained by learning two piece-wise linear curves on the sorted drug era curve. Finally n and *persistent window* were set to $\frac{\tau}{2}$ [21], where τ is the mean of all optimal points of drugs to obtain the final value (Figure 6). In the HbA1c laboratory measurements, n was 31 days but was approximated to 30 days for the ease of clinical interpretablity.

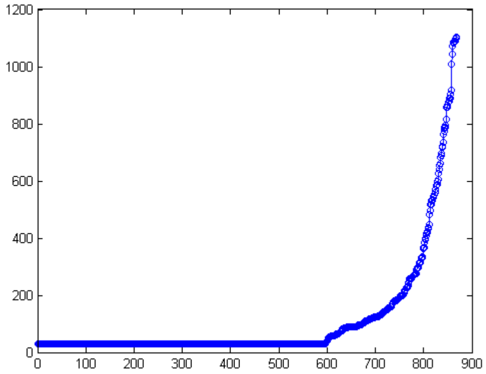


Figure 6: The sorted list of all change points of drug eras of all drugs.

8.2.2 *Building patient similarity network.*

The PerDREP model is generic enough to leverage any patient similarity network using any similarity measure for regularizing the coefficients of different patients. Without loss of generalizability, we consider patients' demographics and diagnosis codes for constructing similarity network. Specifically, we first construct a binary diagnosis vector for each patient containing their ICD9 diagnosis codes until the first occurrence of the HbA1c measurement. Then, we augment this diagnosis vector with demographic variables to compute the similarity measure between two patients. We used two types of similarity measures (cosine and Jaccard) and both produced similar patient similarity networks.

One interesting phenomenon that we observed in the obtained network (Figure 7) is that the similarities among the patients are not uniformly distributed, rather it has multiple modalities, i.e., some patients are close to each other (those lying on the periphery of the sphere in Figure 7), while the rest are very dissimilar to each other (those lying on the center of the sphere). Therefore, we use only the

[1] http://micromedex.com/products/product-suites/clinical-knowledge/redbook
[2] https://www.cdc.gov/nchs/icd/icd9.htm

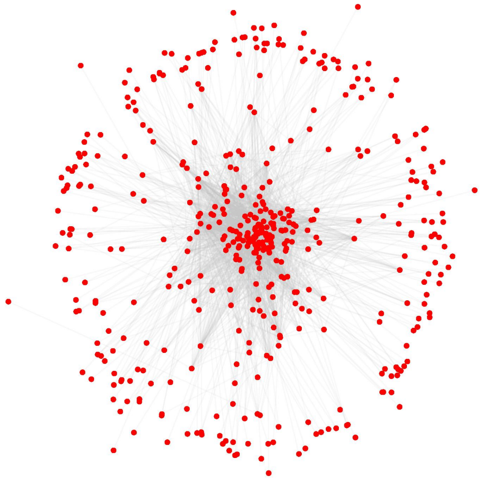


Figure 7: The similarity network constructed out of all patients in the HbA1c cohort.

most similar patients by sparsifying the patient network using a threshold (denoted as *PerDREP-thre*) on the patient similarities. We observed reasonably stable models by choosing the top $N * log(N)$ edges of the network where N is the number of patients. However, this approach often leads to only a few connected components.

Alternatively, the*PerDREP-MST* approach was explored using a sparse fully connected network structure based upon the similarity edges of the patient network. In particular, we constructed a minimum spanning tree (MST) [9] connecting all components of the graph, such that the inverse of similarity scores of the obtained MST is minimized.