Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing

Matthew Fredrikson*, Eric Lantz*, Somesh Jha*, Simon Lin†, David Page*, Thomas Ristenpart*
University of Wisconsin, Marshfield Clinic Research Foundation†*

Abstract

We initiate the study of privacy in pharmacogenetics, wherein machine learning models are used to guide medical treatments based on a patient's genotype and background. Performing an in-depth case study on privacy in personalized warfarin dosing, we show that suggested models carry privacy risks, in particular because attackers can perform what we call *model inversion*: an attacker, given the model and some demographic information about a patient, can predict the patient's genetic markers.

As differential privacy (DP) is an oft-proposed solution for medical settings such as this, we evaluate its effectiveness for building private versions of pharmacogenetic models. We show that *DP mechanisms prevent our model inversion attacks when the privacy budget is carefully selected*. We go on to analyze the impact on utility by performing simulated clinical trials with DP dosing models. We find that for privacy budgets effective at preventing attacks, *patients would be exposed to increased risk of stroke, bleeding events, and mortality*. We conclude that *current* DP mechanisms do not simultaneously improve genomic privacy while retaining desirable clinical efficacy, highlighting the need for new mechanisms that should be evaluated *in situ* using the general methodology introduced by our work.

1 Introduction

In recent years, technical advances have enabled inexpensive, high-fidelity molecular analyses that characterize the genetic make-up of an individual. This has led to widespread interest in *personalized medicine*, which tailors treatments to each individual patient using genotype and other information to improve outcomes. Much of personalized medicine is based on *pharmacogenetic* (sometimes called *pharmacogenomic*) models [3, 14, 21, 40] that are constructed using supervised machine learning over large patient databases containing clinical and genomic data. Prior works [36, 37] in non-medical settings have shown that leaking datasets can enable de-anonymization of users and other privacy risks. In the pharmacogenetic setting, datasets themselves are often only disclosed to researchers, yet the models learned from them are made public (e.g., published in a paper). *Our focus is therefore on determining to what extent the models themselves leak private information, even in the absence of the original dataset.*

To do so, we perform a case study of warfarin dosing, a popular target for pharmacogenetic modeling. Warfarin is an anticoagulant widely used to help prevent strokes in patients suffering from atrial fibrillation (a type of irregular heart beat). However, it is known to exhibit a complex dose-response relationship affected by multiple genetic markers [43], with improper dosing leading to increased risk of stroke or uncontrolled bleeding [41]. As such, a long line of work [3, 14, 16, 21, 40] has sought pharmocogenetic models that can accurately predict proper dosage based on patient clinical history, demographics, and genotype. A review of this literature is given in [23].

Our study uses a dataset collected by the *International Warfarin Pharmocogenetics Consortium* (IWPC), to date the most expansive such database containing demographic information, genetic markers, and clinical histories for thousands of patients from around the world. While this particular dataset is publicly-available in a deidentified form, it is equivalent to data used in other studies that must be kept private (e.g., due to lack of consent to release). We therefore use it as a proxy for a private dataset. The paper authored by IWPC members [21] details methods to learn linear regression models from this dataset, and shows that using the resulting models to predict initial dose outperforms the standard clinical regimen in terms of absolute distance from stable dose. Randomized trials have been done to evaluate clinical effectiveness, but have not yet validated the utility of genetic information [27].

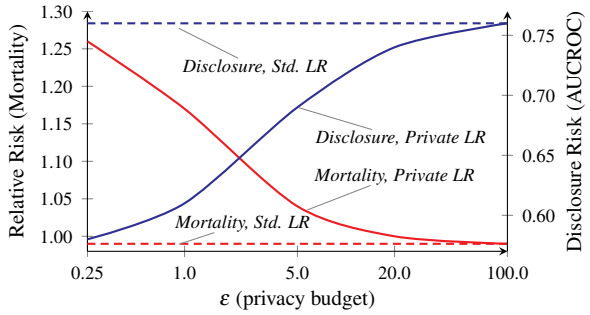

Figure 1: Mortality risk (relative to current clinical practice) for, and VKORC1 genotype disclosure risk of, ε-differentially private linear regression (LR) used for warfarin dosing (over five values of ε, curves are interpolated). Dashed lines correspond to non-private linear regression.

Model inversion. We study the degree to which these models leak sensitive information about patient genotype, which would pose a danger to genomic privacy. To do so, we investigate *model inversion attacks* in which an adversary, given a model trained to predict a specific variable, uses it to make predictions of unintended (sensitive) attributes used as input to the model (i.e., an attack on the privacy of attributes). Such attacks seek to take advantage of correlation between the target, unknown attributes (in our case, demographic information) and the model output (warfarin dosage). A priori it is unclear whether a model contains enough exploitable information about these correlations to mount an inversion attack, and it is easy to come up with examples of models for which attackers will not succeed.

We show, however, that warfarin models do pose a privacy risk (Section 3). To do so, we provide a general model inversion algorithm that is optimal in the sense that it minimizes the attacker's *expected misprediction rate* given the available information. We find that when one knows a target patient's background and stable dosage, their genetic markers are predicted with significantly better accuracy (up to 22% better) than guessing based on marginal distributions. In fact, *it does almost as well as regression models specifically trained to predict these markers (only ~5% worse)*, suggesting that model inversion can be nearly as effective as learning in an "ideal" setting. Lastly, the inverted model performs measurably better for members of the training cohort than others (yielding an increased 4% accuracy) indicating a leak of information specifically about those patients.

Role of differential privacy. Differential privacy (DP) is a popular framework for designing statistical release mechanisms, and is often proposed as a solution to privacy concerns in medical settings [10, 12, 45, 47]. DP is parameterized by a value ε (sometimes referred to as the *privacy budget*), and a DP mechanism guarantees that the likelihood of producing any particular output from an input cannot vary by more than a factor of e^{ε} for "similar" inputs differing in only one subject.

Following this definition in our setting, DP guarantees protection against attempts to infer whether a subject was included in the training set used to derive a machine learning model. It does *not* explicitly aim to protect attribute privacy, which is the target of our model inversion attacks. However, others have motivated or designed DP mechanisms with the goal of ensuring the privacy of patients' diseases [15], features on users' social network profiles [33], and website visits in network traces [38]—all of which relate to attribute privacy. Furthermore, recent theoretical work [24] has shown that in some settings, including certain applications of linear regression, incorporating noise into query results preserves attribute privacy. This led us to ask: can *genomic* privacy benefit from the application of DP mechanisms in our setting?

To answer this question, we performed the first end-to-end evaluation of DP in a medical application (Section 5). We employ two recent algorithms on the IWPC dataset: the *functional mechanism* of Zhang *et al.* [47] for producing private linear regression models, and Vinterbo's *privacy-preserving projected histograms* [44] for producing differentially-private synthetic datasets, over which regression models can be trained. These algorithms represent the current state-of-the-art in DP mechanisms for their respective models, with performance reported by the authors that exceeds previous DP mechanisms designed for similar tasks.

On one end of our evaluation, we apply a model inverter to quantify the amount of information leaked about patient genetic markers by ε-DP versions of the IWPC model. On the other end, we quantify the impact of ε on patient outcomes, performing *simulated clinical trials* via techniques widely used in the medical literature [4, 14, 18, 19]. Our main results, a subset of which are shown in Figure 1, show a clear trade-off between patient outcomes and privacy:

- ***"Small ε"-DP protects genomic privacy***: Even though DP was not specifically designed to protect attribute privacy, we found that for sufficiently small ε (≤ 1), genetic markers cannot be accurately predicted (see the line labeled "Disclosure, private LR" in Figure 1), and there is no discernible difference between the model inverter's performance on the training and validation sets. However, this effect quickly vanishes as ε increases, where genotype is predicted with up to 58% accuracy (0.76 AUCROC). This is significantly (22%) better than the 36% accuracy one achieves without the models, and not far below (5%) the "best possible" performance of a non-private regression model trained to predict the same genotype using IWPC data.

- ***Current DP mechanisms harm clinical efficacy***: Our simulated clinical trials reveal that for $\varepsilon \leq 5$ the risk of fatalities or other negative outcomes increases significantly (up to $1.26\times$) compared to the current clinical practice, which uses non-personalized, fixed dosing and so leaks no information at all. Note that the range of ε (> 5) that provides clinical utility not only fails to protect genomic privacy, but are commonly assumed to provide insufficient DP guarantees as well. (See the line labeled "Mortality, private LR" in Figure 1.)

Put simply: our analysis indicates that in this setting where utility is paramount, the best known mechanisms for our application do not give an ε for which state-of-the-art DP mechanisms can be reasonably employed.

Implications of our results. Our results suggest that there is still much to learn about pharmacogenetic privacy. Differential privacy is suited to settings in which *privacy and utility requirements are not fundamentally at odds*, and can be balanced with an appropriate privacy budget. Although the mechanisms we studied do not properly strike this balance, future mechanisms may be able to do so—the *in situ* methodology given in this paper may help to guide such efforts. In settings where privacy and utility *are* fundamentally at odds, release mechanisms of any kind will fail, and restrictive access control policies may be the best answer. The model inversion techniques outlined here can help to identify these situations, and quantify the risks.

2 Background

Warfarin and Pharmacogenetics Warfarin, also known in the United States by the brand name Coumadin, is a widely prescribed anticoagulant medication. It is used to treat patients suffering from cardiovasvular problems, including atrial fibrillation (a type of irregular heart beat) and heart valve replacement. By reducing the tendency of blood to clot, at appropriate dosages it can reduce risk of clotting events, particularly stroke. Unfortunately, warfarin is also very difficult to dose: proper dosages can differ by an order of magnitude between patients, and this has led to warfarin's status as one of the leading causes of drug-related adverse events in the United States [26]. Underestimating the dose can result in failure to prevent the condition the drug was prescribed to treat. Overestimating the dose can, just as seriously, lead to uncontrolled bleeding events because the drug interferes with clotting. Because of these risks, in existing clinical practice patients starting on warfarin are given a fixed initial dose but then must visit a clinic many times over the first few weeks or months of treatment in order to determine the correct dosage which gives the desired therapeutic effect.

Stable dose is assessed clinically by measuring the time it takes for blood to clot, called prothrombin time. This measure is standardized between different manufacturers as an international normalized ratio (INR). Based on the patient's indication for (i.e., the reason to prescribe) warfarin, a clinician determines a target INR range. After the fixed initial dose, later doses are modified until the patient's INR is within the desired range and maintained at that level. INR in the absence of anticoagulation therapy is approximately 1, while the desired INR for most patients in anticoagulation therapy is in the range 2–3 [5]. INR is the response measured by the physiological model used in our simulations in Section 5.

Genetic variability among patients is known to play an important role in determining the proper dose of warfarin [23]. Polymorphisms in two genes, VKORC1 and CYP2C9, are associated with the mechanism with which the body metabolizes the drug, which in turn affects the dose required to reach a given concentration in the blood. Warfarin works by interfering with the body's ability to recycle vitamin K, which is used to regulate blood coagulation. VKORC1, part of the vitamin K epoxide reductase complex, is a component of the vitamin K cycle. CYP2C9 encodes for a variant of cytochrome P450, a family of proteins which oxidize a variety of medications. Since each person has two copies of each gene, there are several combinations of variants possible. Following the IWPC paper [21], we represent VKORC1 polymorphisms by single nucleotide polymorphism (SNP) rs9923231, which is either G (common variant) or A (uncommon variant), resulting in three combinations G/G, A/G, or A/A. Similarly, CYP2C9 variants are *1 (most common), *2, or *3, resulting in 6 combinations.

Taken together with age and height, Sconce *et al.* [40] demonstrated that CYP2C9 and VKORC1 account for 54% of the total warfarin dose requirement variability. In turn, a large literature (over 50 papers as of early 2013) has sought pharmacogenetic algorithms that predict proper dose by taking advantage of patient genetic markers for CYP2C9 and VKORC1, together with demographic information and clinical history (e.g., current medications). These typically involve learning a simple predictive model of stable dose from previously obtained outcomes. We focus on the IWPC algorithm [21], a study resulting in production of a linear regression model that, when used to predict the initial dosage, has been shown to provide improved outcomes in simulated clinical trials using the IWPC dataset discussed below. Interestingly, linear regression performed as well or better than a wide variety of other, more complex machine learning techniques. Some pharmacogenetic algorithms for warfarin are currently also undergoing (real) clinical trials [1].

Dataset The IWPC [21] collected data on patients who were prescribed warfarin from 21 sites in 9 countries on 4 continents. The data was curated by staff at the Pharmacogenomics Knowledge Base [2], and each site obtained informed consent to use de-identified data from patients prior to the study. Because the dataset contains no protected health information, and the Pharmacogenomics Knowledge Base has since made the dataset publicly available for research purposes, it is exempt from institutional review board review. However, the type of data contained in the IWPC dataset is equivalent to many other medical datasets that have not been released publicly [3, 7, 16, 40], and are considered private.

Each patient was genotyped for at least one SNP in VKORC1, and for variants of CYP2C9. In addition, other information such as age, height, weight, race, and other medications was collected. The outcome variable is the stable therapeutic dose of warfarin, defined as the steady-state dose that led to stable anticoagulation levels. The patients in our dataset were restricted to those with target INR in the range 2–3 (the vast majority of patients), as is standard practice with most studies of warfarin dosing efficacy [3, 14]. We divided the data into two cohorts based on those used in IWPC [21]. The first (training) cohort was used to build a set of pharmacogenetic dosing algorithms. The second (validation) cohort was used to test privacy attacks as well as draw samples for the clinical simulations. To make the data suitable for regression we removed all patients missing CYP2C9 or VKORC1 genotype, normalized the data to the range [-1,1], converted all nominal attributes into binary-valued numeric attributes, and scaled each row into the unit sphere. Our eventual training cohort consisted of 2644 patients, and our validation cohort of 853 patients, and corresponds to the same training-validation split used by IWPC (but without the missing values used in the IWPC split).

3 Privacy of Pharmacogenetic Models

In this section we investigate the risks involved in releasing regression models trained over private data, using models that predict warfarin dose as our case study. We consider a setting where an adversary is given access to such a model, the warfarin dosage of an individual, some rudimentary information about the data set, and possibly some additional attributes about that individual. The adversary's goal is to predict one of the *genotype* attributes for that individual. In order for this setting to make sense, the genotype attributes, warfarin dose, and other attributes known to the adversary must all have been in the private data set. We emphasize that the techniques introduced can be applied more generally, and save as future work investigating other pharmacogenetic settings.

3.1 Attack Model

We assume an adversary who employs an inference algorithm $\mathcal{A}$ to discover the genotype (in our experiments, either CYP2C9 or VKORC1) of a target individual α. The adversary has access to a linear model f trained over a dataset D drawn i.i.d. from an unknown prior distribution p. D has domain $\mathbf{X} \times Y$, where $\mathbf{X} = X_1, \ldots, X_d$ is the domain of possible *attributes* and Y is the domain of the *response*. α is represented by a single row in D, $(\mathbf{x}^\alpha, y^\alpha)$, and the attribute learned by the adversary is referred to as the *target attribute* $\mathbf{x}_t^\alpha$.

In addition to f, the adversary has access to marginals[1] $p_{1,\ldots,d,y}$ of the joint prior p, the dataset domain $\mathbf{X} \times Y$, α's stable dosage y^α of warfrain, some information π about f's performance (details in the following section), and either of the following subsets $\mathbf{x}_K^\alpha$ of α's attributes:

- *Basic demographics*: a subset of α's demographic data, including age (binned into eight groups by the IWPC), race, height, and weight (denoted $x_{\text{age}}^\alpha, x_{\text{race}}^\alpha, \ldots$). Note that this corresponds to a subset of the non-genetic attributes in D.
- *All background*: all of p's attributes except CYP2C9 or VKORC1 genotype.

The adversary has black-box access to f. Unless it is clear from the context, we will specify whether f is the output of a DP mechanism, and which type of background information is available.

3.2 Model Inversion

In this section, we discuss a technique for inferring CYP2C9 and VKORC1 genotype from a model designed to predict warfarin dosing. Given a model f that takes inputs $\mathbf{x}$ and outputs a predicted stable dose y, the attacker seeks to build an algorithm $\mathcal{A}$ that takes as input some subset $\mathbf{x}_K^\alpha$ of attributes (corresponding to demographic or additional background attributes from $\mathbf{X}$), a known stable dose y^α, and outputs a prediction of $\mathbf{x}_t$ (corresponding either to CYP2C9 or VKORC1). We begin by presenting a general-purpose algorithm, and show how it can be applied to linear regression models.

A general algorithm. We present an algorithm for model inversion that is independent of the underlying model structure (Figure 2). The algorithm works by estimating the probability of a potential target attribute given the available information and the model. Its operation is straightforward: *candidate* database rows that are similar to what is known about α are run *forward* through

[1] These are commonly published in studies, and when it is clear from the context, we will drop the subscript.

1. Input: $\mathbf{z}_K = (x_1, \ldots, x_k, y), f, p_{1,\ldots,d,y}$
2. Find the *feasible set* $\hat{\mathbf{X}} \subseteq \mathbf{X}$, i.e., such that $\forall \mathbf{x} \in \hat{\mathbf{X}}$
   (a) $\mathbf{x}$ matches $\mathbf{z}_K$ on known attributes: for $1 \leq i \leq k, \mathbf{x}_i = x_i$.
   (b) f evaluates to y as given in $\mathbf{z}_K$: $f(\mathbf{x}) = y$.
3. If $|\hat{\mathbf{X}}| = 0$, return $\perp$.
4. Return x_t that maximizes $\sum_{\mathbf{x} \in \hat{\mathbf{X}}: \mathbf{x}_t = x_t} \prod_{1 \leq i \leq d} p_i(\mathbf{x}_i)$

(a) $\mathcal{A}_0$: Model inversion without performance statistics.

1. Input: $\mathbf{z}_K = (x_1, \ldots, x_k, y)$, f, π, $p_{1,\ldots,d,y}$
2. Find the *feasible set* $\hat{\mathbf{X}} \subseteq \mathbf{X}$, i.e., such that $\forall \mathbf{x} \in \hat{\mathbf{X}}$
   (a) $\mathbf{x}$ matches $\mathbf{z}_K$ on known attributes: for $1 \leq i \leq k$, $\mathbf{x}_i = x_i$.
3. If $|\hat{\mathbf{X}}| = 0$, return $\perp$.
4. Return x_t that maximizes $\sum_{\mathbf{x} \in \hat{\mathbf{X}}: \mathbf{x}_t = x_t} \pi_{y, f(\mathbf{x})} \prod_{1 \leq i \leq d} p_i(\mathbf{x}_i)$

(b) $\mathcal{A}_\pi$: Model inversion with performance statistics π.

Figure 2: Model inversion algorithm.

the model. Based on the known priors, and how well the model's output on that row coincides with α's known response value, the candidate rows are weighted. The target attribute with the greatest weight, computed by marginalizing the other attributes, is returned.

Below, we describe this algorithm in more detail. We derive each step by showing how to compute the *least biased* estimate of the target attribute's likelihood, which the model inversion algorithm maximizes to form a prediction. As we reason below, this approach is *optimal* in the sense that it minimizes the expected misclassification rate when the adversary has no other information (i.e., makes no further assumptions) beyond what is given in Section 3.1.

Derivation. We begin the description with a simpler restricted case in which the model always produces the correct response. Assume for now that f is *perfect*, i.e., it never makes a misprediction, and we can assume that $f(\mathbf{x}) = y$ almost surely for any sample $(\mathbf{x}, y)$; this case is covered by $\mathcal{A}_0$ in Figure 2. In the following, we assume the sample corresponds to the individual α, and drop the superscript for clarity. Suppose the adversary wishes to learn the probability that $\mathbf{x}_t$ takes a certain value x_t, i.e., $\mathbf{Pr}[\mathbf{x}_t = x_t | \mathbf{x}_K, y]$, given some known attributes $\mathbf{x}_K$, response variable y, and the model f. Here, and in the following discussion, the probabilities in $\mathbf{Pr}[\cdot]$ expressions are always over draws from the unknown joint prior p unless stated otherwise. Let $\hat{\mathbf{X}} = \{\mathbf{x}' : \mathbf{x}'_K = \mathbf{x}_K \text{ and } f(\mathbf{x}') = y\}$ be the subset of $\mathbf{X}$ matching the given information $\mathbf{x}_K$ and y. Then by straightforward computation,

$$\mathbf{Pr}[x_t | \mathbf{x}_K, y] = \frac{\mathbf{Pr}[x_t, \mathbf{x}_K, y]}{\mathbf{Pr}[\mathbf{x}_K, y]} = \frac{\sum_{\mathbf{x}' \in \hat{\mathbf{X}}: \mathbf{x}'_t = x_t} p(\mathbf{x}', y)}{\sum_{\mathbf{x}' \in \hat{\mathbf{X}}} p(\mathbf{x}', y)} \quad (1)$$

Now, the adversary does not know the true underlying joint prior p. He only knows the marginals $p_{1,\ldots,d,y}$, so any distribution with these marginals is a possible prior. To characterize the unbiased prior that satisfies these constraints, we apply the *prinicipal of maximum entropy*[2] [22], which in our setting gives the prior:

$$p(\mathbf{x}, y) = p(y) \cdot \prod_{1 \leq i \leq d} p(\mathbf{x}_i) \quad (2)$$

Continuing with the previous expression, we now have,

$$\mathbf{Pr}[x_t | \mathbf{x}_K, y] = \frac{\sum_{\mathbf{x}' \in \hat{\mathbf{X}}: \mathbf{x}'_t = x_t} p(y) \prod_i p(\mathbf{x}'_i)}{\sum_{\mathbf{x}' \in \hat{\mathbf{X}}} p(y) \prod_i p(\mathbf{x}'_i)} \quad (3)$$

$$\propto \sum_{\mathbf{x}' \in \hat{\mathbf{X}}: \mathbf{x}'_t = x_t} \prod_i p(\mathbf{x}'_i) \quad (4)$$

This last step follows because the denominator is independent of the choice of x_t. Notice that this is exactly the quantity that is maximized by the value returned by $\mathcal{A}_0$ (Figure 2 (a)). This is the *maximum a posteriori probability* (MAP) estimate, which minimizes the adversary's expected misclassification rate. Under these assumptions, $\mathcal{A}_0$ is an optimal algorithm for model inversion.

$\mathcal{A}_\pi$ in Figure 2 (b) generalizes this reasoning to the case where f is not assumed to be perfect, and the adversary has information about the performance of f over samples drawn from p. We model this information with a function π, defined in terms of a random sample $\mathbf{z}$ from p,

$$\pi(y, y') = \mathbf{Pr}\left[\mathbf{z}_y = y | f(\mathbf{z}_\mathbf{x}) = y'\right] \quad (5)$$

In other words, $\pi(y, y')$ gives the probability that the true response drawn with attributes $\mathbf{z}_\mathbf{x}$ is y given that the model outputs y'. We write $\pi_{y,y'}$ to simplify notation. In practice, π can be estimated using statistics commonly released with models, such as confusion matrices or standardized regression error.

Because f is not assumed to be perfect in the general setting, $\hat{\mathbf{X}}$ is defined slightly differently than in $\mathcal{A}_0$; the second restriction, that $f(\mathbf{x}^\alpha) = y^\alpha$, is removed. After constructing $\hat{\mathbf{X}}$, $\mathcal{A}_\pi$ uses the marginals and π to weight each candidate $\mathbf{x} \in \hat{\mathbf{X}}$ by the probability that f behaves as observed (i.e., outputs $f(\mathbf{x})$) when the response variable matches what the adversary knows to be true (i.e.,

[2] *cf.* Jaynes [22], "[The maximum entropy prior] is least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information."

y). Again, using the maximum entropy prior from before gives the MAP estimate in the more general setting,

$$\mathbf{Pr}[x_t|\mathbf{x}_K, y^\alpha, f] = \frac{\sum_{\mathbf{x}'\in\hat{\mathbf{X}}:\mathbf{x}'_t=x_t} \mathbf{Pr}[\mathbf{x}', y, f(\mathbf{x}')]}{\sum_{\mathbf{x}'\in\hat{\mathbf{X}}} \mathbf{Pr}[\mathbf{x}', y, f(\mathbf{x}')]} \quad (6)$$

$$= \frac{\sum_{\mathbf{x}'\in\hat{\mathbf{X}}:\mathbf{x}'_t=x_t} \mathbf{Pr}[y|\mathbf{x}', f(\mathbf{x}')]\, p(\mathbf{x}')}{\sum_{\mathbf{x}'\in\hat{\mathbf{X}}} \mathbf{Pr}[\mathbf{x}', y, f(\mathbf{x}')]} \quad (7)$$

$$\propto \sum_{\mathbf{x}'\in\hat{\mathbf{X}}:\mathbf{x}'_t=x_t} \pi_{y,f(\mathbf{x}')} \left(\prod_i p(\mathbf{x}'_i)\right) \quad (8)$$

The second step follows from the independence of the maximum entropy prior in our setting, and the fact that $\mathbf{x}$ determines $f(\mathbf{x})$ so $\mathbf{Pr}[f(\mathbf{x}'), \mathbf{x}'] = \mathbf{Pr}[\mathbf{x}']$.

Application to linear regression. Recall that a linear regression model assumes that the response is a linear function of the attributes, i.e., there exists a coefficient vector $\mathbf{w} \in \mathbb{R}^d$ and random *residual error* δ such that $y = \mathbf{w}^T\mathbf{x} + b + \delta$ for some bias term b. A linear regression model f_L is then an estimate $(\hat{\mathbf{w}}, \hat{b})$ of $\mathbf{w}$ and the bias term, which operates as: $f_L(\mathbf{x}) = \hat{b} + \hat{\mathbf{w}}^T\mathbf{x}$. It is typical to assume that δ has a fixed Gaussian distribution $\mathcal{N}(0, \sigma^2)$ for some variance σ. Most regression software estimates σ^2 empirically from training data, so it is often published alongside a linear regression model. Using this the adversary can derive an estimate of π,

$$\hat{\pi}(y, y') = \mathbf{Pr}_{\mathcal{N}(0,\sigma^2)}[y - y']$$

Steps 2 and 4 of $\mathcal{A}_\pi$ may be expensive to compute if $|\hat{\mathbf{X}}|$ is large. In this case, one can approximate using Monte Carlo techniques to sample members of $\hat{\mathbf{X}}$. Fortunately, in our setting, the nominal-valued variables all come from sets with small cardinality. The continuous variables have natural discretizations, as they correspond to attributes such as age and weight. Thus, step 4 can be computed directly by taking a discrete convolution over the unknown attributes without resorting to approximation.

Discussion. We have argued that $\mathcal{A}_\pi$ is optimal in one particular sense, i.e., it minimizes the expected misclassification rate on the maximum-entropy prior given the available information (the model and marginals). However, it is not hard to specify joint priors p for which the marginals $p_{1,\ldots,d,y}$ convey little useful information, so the expected misclassification rate minimized here diverges substantially from the true rate. In these cases, $\mathcal{A}_\pi$ may perform poorly, and more background information is needed to accurately predict model inputs.

There is also the possibility that the model itself does not contain enough useful information about the correlation between certain input attributes and the output. For illustrative purposes, consider a model taking one input attribute, that discards all information about that attribute except a single bit, e.g., it performs a comparison with a fixed constant. If the attribute is distributed uniformly across a large domain, then $\mathcal{A}_\pi$ will only perform negligibly better than guessing from the marginal. Thus, determining how well a model allows one to predict sensitive inputs generally requires further analysis, which is the purpose of the evaluation that we discuss next (see also Section 4).

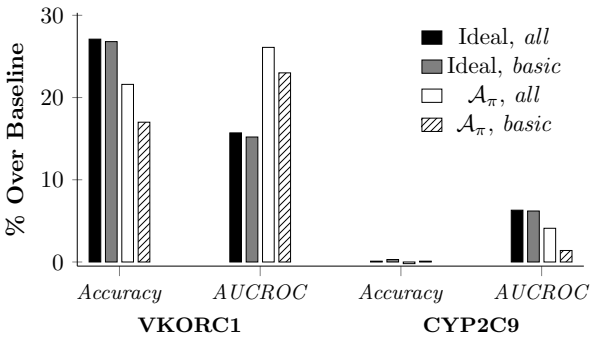

Figure 3: Model inversion performance, as improvement over baseline guessing from marginals, given a linear model derived from the training data. Available background information specified by *all* and *basic* as discussed in Section 3.1.

Results on non-private regression. To evaluate $\mathcal{A}_\pi$, we split the IWPC dataset into a training and validation set (see Section 2), D_T and D_V respectively, use D_T to derive a least-squares linear model f, and then run $\mathcal{A}_\pi$ on every α in D_T with either of the two background information types (*all* or *basic*, see Section 3.1) to predict both genotypes. In order to determine how how well one can predict these genotypes in an *ideal* setting, we built and evaluated a multinomial logistic regression model (using R's `nnet` package) for each genotype from the IWPC data. This allows us to compare the performance of $\mathcal{A}_\pi$ against "best-possible" results achieved using standard machine learning techniques with linear models.

We measure performance both in terms of *accuracy*, which is the percentage of samples for which the algorithm correctly predicted genotype, and *AUCROC*, which is the multi-class area under the ROC curve defined by Hand and Till [17]. While accuracy is generally easier to interpret, it can give a misleading characterization of predictive ability for *skewed* distributions—if the predicted attribute takes a particular value in 75% of the samples, then a trivial algorithm can easily obtain 75% accuracy by always guessing this value. AUCROC does not suffer this limitation, and so gives a more balanced characterization of how well an algorithm predicts both common and rare values.

The results are given in Figure 3, which shows the performance of $\mathcal{A}_\pi$ and "ideal" multinomial regression predicting VKORC1 and CYP2C9 on the training set. The numbers are given relative to the baseline performance obtained by always guessing the most probable genotype based on the given marginal prior–36% accuracy on VKORC1, 75% accuracy on CYP2C9, and 0.5 AUCROC for both genotypes. We see that $\mathcal{A}_\pi$ comes close to ideal accuracy on VKORC1 (5% less accurate with all background information), and actually exceeds the ideal predictor in terms of AUCROC. This means that $\mathcal{A}_\pi$ does a better job predicting *rare* genotypes than the ideal model, but does slightly worse overall, and may be a result of the ideal model avoiding overfitting to uncommon data points.

The results for CYP2C9 are quite different. Neither $\mathcal{A}_\pi$ or the ideal model were able to predict this genotype more accurately than baseline. This indicates that CYP2C9 is difficult to predict using linear models, and because we use a linear model to run $\mathcal{A}_\pi$ in this case, it is no surprise that it inherits this limitation. Both the ideal model and $\mathcal{A}_\pi$ slightly outperform baseline prediction in terms of AUCROC, and $\mathcal{A}_\pi$ comes very close to ideal performance (within 2%). In one case $\mathcal{A}_\pi$ does slightly worse (0.2%) than baseline accuracy; this may be due to the fact that the marginals and $\hat{\pi}$ used by $\mathcal{A}_\pi$ are approximations to the true marginals and error distribution π.

We also evaluated $\mathcal{A}_\pi$ on the validation set (using a model f derived from the training set). We found that both genotypes are predicted more accurately on the training set than validation. For VKORC1, $\mathcal{A}_\pi$ was 3% more accurate and yielded an additional 4% AUCROC. The difference was less pronounced with CYP2C9, which was 1.5% more accurate with an additional 2% AUCROC. Although these differences are not as large as the aboslute gain over baseline prediction, they persist across other training/validation splits. We ran 100 instances of cross-validation, and measured the difference between training and validation performance. We found that we were on average able to better predict the training cohort ($p < 0.01$).

4 Differentially-Private Mechanisms and Pharmacogenetics

In the last section, we saw that linear models trained on private datasets leak information about patients in the training cohort. In this section, we explore the issue on models and datasets for which differential privacy has been applied.

As in the previous section, we take the perspective of the adversary, and attempt to infer patients' genotype given differentially-private models and different types of background information on the targeted individual. As such, we use the same attack model, but rather than assuming the adversary has access to f, we assume access to a *differentially private version of* the original dataset D or f. We use two published differentially-private mechanisms with publicly-available implementations: *private projected histograms* [44] and the *functional mechanism* [47] for learning private linear regression models. Although full histograms are typically not published in pharmacogenetic studies, we analyze their privacy properties here to better understand the behavior of differential privacy across algorithms that implement it differently.

Our key findings are summarized as follows:

- Some ε values effectively protect genomic privacy for DP linear regression. For $\varepsilon \leq 1$, $\mathcal{A}_\pi$ could not predict VKORC1 better on the training set than the validation set either in terms of accuracy or AUCROC. The same result holds on CYP2C9, but only when measured in terms of AUCROC. $\mathcal{A}_\pi$'s *absolute* performance for these ε is not much better than the baseline either: VKORC1 is predicted only 5% better at $\varepsilon = 1$, and CYP2C9 sees almost no improvement.

- "Large"-ε DP mechanisms offer little genomic privacy. When $\varepsilon \geq 5$, both DP mechanisms see a statistically-significant increase in training set performance over validation ($p < 0.02$), and as ε approaches 20 there is little difference from non-private mechanisms (between 3%-5%).

- Private histograms disclose significantly more information about genotype than private linear regression, even at identical ε values. At all tested ε, private histograms leaked more on the training than validation set. *This result holds even for non-private regression models*, where the AUCROC gap reached 3.7% area under the curve, versus the 3.9% - 5.9% gap for private histograms. This demonstrates that the relative nature of differential privacy's guarantee can lead to meaningful concerns.

Our results indicate that understanding the implications of differential privacy for pharmacogenomic dosing is a difficult matter—even small values of ε might lead to unwanted disclosure in many cases.

Differential Privacy Dwork introduced the notion of differential privacy [11] as a constructive response to an impossibility result concerning stronger notions of private data release. For our purposes, a dataset D is a number m of vector, value pairs $(\mathbf{x}^{\alpha_1}, y^{\alpha_1}), \ldots, (\mathbf{x}^{\alpha_m}, y^{\alpha_m})$

where $\alpha_1, \ldots, \alpha_m$ are (randomized) patient identifiers, each $\mathbf{x}^{\alpha_i} = [x_1^{\alpha_i}, \ldots, x_d^{\alpha_i}]$ is a patient's demographic information, age, genetic variants, etc., and y^{α_i} is the stable dose for patient α_i. A (differential) privacy mechanism K is a randomized algorithm that takes as input a dataset D and, in the cases we consider, either outputs a new dataset D_{priv} or a linear model M_{priv} (i.e., a real- valued linear function with n inputs. We denote the set of possible outputs of a mechanism as Range(K).

A mechanism K achieves ε-differential privacy if for all databases D_1, D_2 differing in at most one row, and all $S \subseteq \text{Range}(\mathsf{K})$,

$$\Pr[\mathsf{K}(D_1) \in S] \leq \exp(\varepsilon) \times \Pr[\mathsf{K}(D_2) \in S]$$

Differential privacy is an information-theoretic guarantee, and holds regardless of the auxiliary information an adversary posesses about the database.

Differentially-private histograms. We first investigate a mechanism for creating a differentially-private version of a dataset via the private projected histogram method [44]. DP datasets are appealing because an (untrusted) analyst can operate with more freedom when building a model; he is free to select whichever algorithm or representation best suits his task, and need not worry about finding differentially-private versions of the best algorithms.

Because the numeric attributes in our dataset are too fine-grained for effective histogram computation, we first discretize each numeric attribute into equal-width bins. In order to select the number of bins, we use a heuristic given by Lei [32] and suggested by Vinterbo [44], which says that when numeric attributes are scaled to the interval $[0,1]$, the bin width is given by $(\log(n)/n)^{1/(d+1)}$, where $n = |D|$ and d is the dimension of D. In our case, this is implies two bins for each numeric attribute. We validated this parameter against our dataset by constructing 100 differentially-private datasets at $\varepsilon = 1$ with $2, 3, 4$, and 5 bins for each numeric attribute, and measured the accuracy of a dose-predicting linear regression model over each dataset. The best accuracy was given for $k = 2$, with the difference in means for $k = 2$ and $k = 3$ not attributable to noise. When the discretized attributes are translated into a private version of the original dataset, the median value from each bin is used to create numeric values.

To infer the private genomic attributes given a differentially-private version D_ε of a dataset, we can compute an empirical approximation $\hat{p}$ to the joint probability distribution p (see Section 3.1) by counting the frequency of tuples in D_ε. A minor complication arises due to the fact that numeric values in D_ε have been discretized and re-generated from the median of each bin. Therefore, the likelihood of finding a row in D_ε that matches any row in D_T or D_V is low. To account for this, we transform each numeric attribute in the background information to the nearest median from the corresponding attribute used in the discretization step when generating D_ε. We then use $\hat{p}$ to directly compute a prediction of the genotype x_t that maximizes $\Pr_{\hat{p}}[\mathbf{x}_t^\alpha = x_t | \mathbf{x}_K^\alpha, y^\alpha]$.

Differentially-private linear regression. We also investigate use of the functional mechanism [47] for producing differentially-private linear regression models. The functional mechanism works by adding Laplacian noise to the coefficients of the objective function used to drive linear regression. This technique stands in contrast to the more obvious approach of directly perturbing the output coefficients of the regression training algorithm, which would require an explicit sensitivity analysis of the training algorithm itself. Instead, deriving a bound on the amount of noise needed for the functional mechanism involves a fairly simple calculation on the objective function [47].

We produce private regression models on the IWPC dataset by first projecting the columns of the dataset into the interval $[-1, 1]$, and then scaling the non-response columns (i.e., all those except the patient's dose) of each row into the unit sphere. This is described in the paper [47] and performed in the publicly-available implementation of the technique, and is necessary to ensure that sufficient noise is added to the objective function (i.e., the amount of noise needed is not scale-invariant). In order to inter-operate with the other components of our evaluation apparatus, we re-implemented the algorithm in R by direct translation from the authors' Matlab implementation. We evaluated the accuracy of our implementation against theirs, and found no statistically-significant difference.

Applying model inversion to the functional mechanism is straightforward, as our technique from Section 3.2 makes no assumptions about the internal structure of the regression model or how it was derived. However, care must be taken with regards to data scaling, as the functional mechanism classifier is trained on scaled data. When calculating $\hat{X}$, all input variables must be transformed by the same scaling function used on the training data, and the predicted response must be transformed by the inverse of this function.

Results on private models. We evaluated our inference algorithms on both mechanisms discussed above at a range of ε values: 0.25, 1, 5, 20, and 100. For each algorithm and ε, we generated 100 private models on the training cohort, and attempted to infer VKORC1 and CYP2C9 for each individual in both the training and validation cohort. All computations were performed in R.

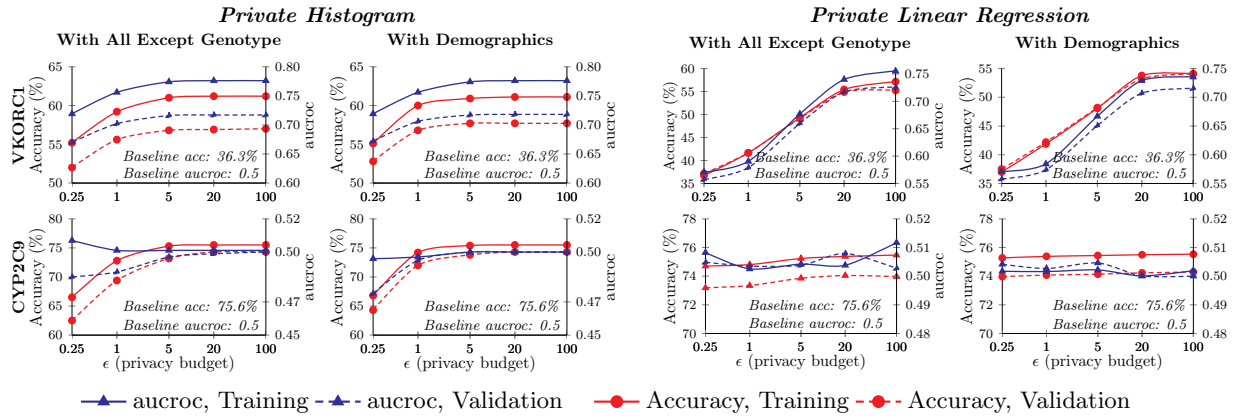

Figure 4: Inference performance for genomic attributes over IWPC training and validation set for private histograms (**left**) and private linear regression (**right**), assuming both configurations for background information. Dashed lines represent accuracy, solid lines area under the ROC curve (AUCROC).

Figure 4 shows our results in detail. In the following, we discuss the main takeaway points.

Private Histograms vs. Linear Regression. We found that private histograms leaked significantly more information about patient genotype than private linear regression models. The difference in AUCROC for histograms versus regression models is statistically significant for VKORC1 at all ε. As Figure 4 indicates, the magnitude of the difference from baseline is also higher for histograms when considering VKORC1, nearly reaching 0.8 AUCROC and 63% accuracy, while regression models achieved at most 0.75 AUCROC and 55–60% accuracy. The AUCROC performance for VKORC1 was greater than the baseline for all ε. However, for CYP2C9 this result only held when assuming all background information except genotype, and only for $\varepsilon \leq 5$; when we assumed only demographic information, there was no significant difference between baseline and private histogram performance.

Disclosure from Overfitting. In nearly all cases, we were able to better infer genotype for patients in the training set than those in the validation set. For private linear regression, this result holds for VKORC1 at $\varepsilon \geq 5.0$ for AUCROC. This is not an artifact of the training/validation split chosen by the IWPC; we ran 10-fold cross validation 100 times, measuring the AUCROC difference between training and test set validation, and found a similar difference between training and validation set performance ($p < 0.01$). The fact that the difference at certain ε values is not statistically significant is evidence that private linear regression is effective at preventing genotype disclosure at these ε. For private histograms, this result held for VKORC1 at all ε, and CYP2C9 at $\varepsilon < 5$ with all background information but genotype.

Differences in Genotype. For both private regression and histogram models, performance for CYP2C9 is strikingly lower than for VKORC1. Private regression models performed no differently from the baseline, achieving essentialy no gain in terms of accuracy and at most 1% gain in AUCROC. We observe that this also held in the non-private setting, and the ideal model achieved the same accuracy as baseline, and only 7% greater AUCROC. This indicates that CYP2C9 is not well-predicted using linear models, and $\mathcal{A}_\pi$ performed nearly as well as is possible.

5 The Cost of Privacy: Negative Outcomes

In addition to privacy, we are also concerned with the utility of a warfarin dosing model. The typical approach to measuring this is a simple accuracy comparison against known stable doses, but ultimately we're interested in how errors in the model will affect patient health. In this section, we evaluate the potential medical consequences of using a differentially-private regression algorithm to make dosing decisions in warfarin therapy. Specifically, we estimate the increased risk of stroke, bleeding, and fatality resulting from the use of differentially-private warfarin dosing at several privacy budget settings. This approach differs from the normal methodology used for evaluating the utility of differentially-private data mining techniques. Whereas evaluation typically ends with a comparison of simple predictive accuracy against non-private methods, we actually simulate the application of a privacy-preserving technique to its domain-specific task, and compare the

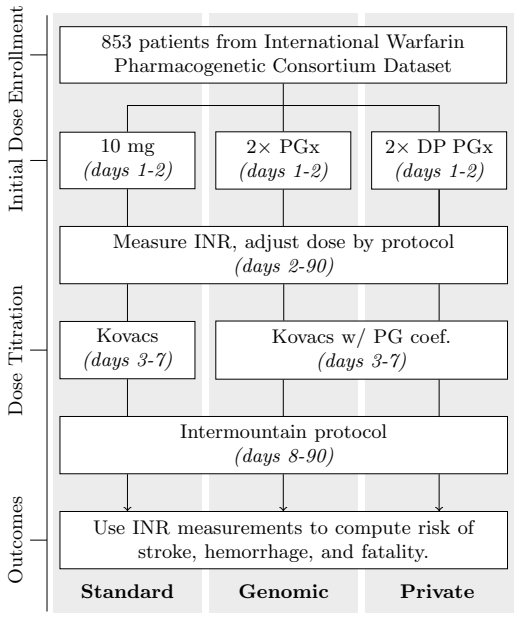

Figure 5: Overview of the Clinical Trial Simulation. *PGx* signifies the pharmacogenomic dosing algorithm, and *DP* differential privacy. The trial consists of three arms differing primarily on initial dosing strategy, and proceeds for 90 days. Details of Kovacs and Intermountain protocol are given in Section 5.3.

outcomes of that task to those achieved without the use of private mechanisms.

5.1 Overview

In order to evaluate the consequences of private genomic dosing algorithms, we simulate a clinical trial designed to measure the effectiveness of new medication regimens. The practice of simulating clinical trials is well-known in the medical research literature [4, 14, 18, 19], where it is used to estimate the impact of various decisions before initiating a costly real-world trial involving human subjects. Our clinical trial simulation follows the design of the CoumaGen clinical trials for evaluating the efficacy of pharmacogenomic warfarin dosing algorithms [3], which is the largest completed real-world clinical trial to date for evaluating these algorithms. At a high level, we train a pharmacogenomic warfarin dosing algorithm and a set of private pharmacogenomic dosing algorithms on the training set. The simulated trial draws random patient samples from the validation set, and for each patient, applies three dosing algorithms to determine the simulated patient's starting dose: the current standard clinical algorithm, the non-private pharmacogenomic algorithm, and one of the private pharmacogenomic algorithms. We then simulate the patient's physiological respose to the doses given by each algorithm using a dose *titration* (i.e., modification) protocol defined by the original CoumaGen trial.

In more detail, our trial simulation defines three parallel *arms* (see Figure 5), each corresponding to a distinct method for assigning the patient's initial dose of warfarin:

1. *Standard:* the current standard practice of initially prescribing a fixed 10mg/day dose.
2. *Genomic:* Use of a genomic algorithm to assign the initial dose.
3. *Private:* Use of a differentially-private genomic algorithm to assign initial dose.

Within each arm, the trial proceeds for 90 simulated days in several stages, as depicted in Figure 5:

1. *Enrollment:* A patient is sampled from the population distribution, and their genotype and demographic characteristics are used to construct an instance of a *Pharmacokinetic/Pharmacodynamic (PK/PD) Model* that characterizes relevant aspects of their physiological response to warfarin (i.e., INR). The PK/PD model contains random variables that are parameterized by genotype and demographic information, and are designed to capture the variance observed in previous population-wide studies of physiological response to warfarin [16].
2. *Initial Dosing:* Depending on which arm of the trial the current patient is in, an initial dose of warfarin is prescribed and administered for the first two days of the trial.
3. *Dose Titration:* For the remaining 88 days of the simulated trial, the patient administers a prescribed dose every 24 hours. At regular intervals specified by the titration protocol, the patient makes "clinic visits" where INR response to previous doses is measured, a new dose is prescribed based on the measured response, and the next clinic visit is scheduled based on the patient's INR and current dose. This is explained in more detail in Sections 5.3 and 5.4.
4. *Measure Outcomes:* The measured responses for each patient at each clinic visit are tabulated, and the risk of negative outcomes is computed.

5.2 Pharmacogenomic Warfarin Dosing

To build the non-private regression model, we use regularized least-squares regression in R, and obtained 15.9% average absolute error (see Figure 6). To build

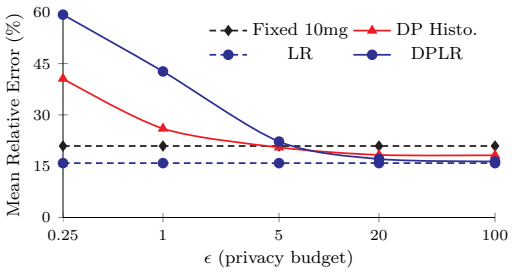

Figure 6: Pharmacogenomic warfarin dosing algorithm performance measured against clinically-deduced ground truth in IWPC dataset.

differentially-private models, we use two techniques: the functional mechanism of Zhang *et al.* [47] and regression models trained on Vinterbo's private projected histograms [44].

To obtain a baseline estimate of these algorithms' performance, we constructed a set of regression models for various privacy budget settings ($\varepsilon = 0.25, 1, 5, 20, 100$) using each of the above methods. The average absolute predictive error, over 100 distinct models at each parameter level, is shown in Figure 6. Although the average error of the private algorithms at low privacy budget settings is quite high, it is not clear how that will affect our simulated patients. In addition to the magnitude of the error, its *direction* (i.e., whether it under- or over-prescribes) matters for different types of risk. Futhermore, because the patient's initial dose is subsequently titrated to more appropriate values according to their INR response, it may be the case that a poor guess for initial dose, as long as the error is not too significant, will only pose a risk during the early portion of the patient's therapy, and a negligible risk overall. Lastly, the accuracy of the standard clinical and non-private pharmacogenomic algorithms are moderate (~15% and 21% error, respectively), and these are the best known methods for predicting initial dose. The difference in accuracy between these and the private algorithm is not extreme (e.g., greater than an order of magnitude), so lacking further information about the correlation between initial dose accuracy and patient outcomes, it is necessary to study their use in greater detail. Removing this uncertainty is the goal of our simulation-based evaluation.

5.3 Dose Assignment and Titration

To assign initial doses and control the titration process in our simulation, we follow the protocol used by the CoumaGen clinical trials on pharmacogenomic warfarin dosing algorithms [3]. In the standard arm, patients are given 10-mg doses on days 1 and 2, followed by dose adjustment according to the Kovacs protocol [29] for days 3 to 7, and final adjustment according to the Intermountain Healthcare protocol [3] for days 8 to 90. Both the Kovacs and Intermountain protocols assign a dose and next appointment time based on the patient's current INR, and possibly their previous dose.

The genomic arm differs from the standard arm for days 1-7. The initial dose for days 1-2 is predicted by the pharmacogenomic regression model, and multiplied by 2 [3]. On days 3-7, the Kovacs protocol is used, but the prescribed dose is multiplied by a coefficient C_{pg} that measures the ratio of the predicted pharmacogenomic dose to the standard 10mg initial dose: $C_{\text{pg}} = (\textit{Initial Pharmacogenomic Dose})/(5\text{ mg})$. On days 8-90, the genomic arm proceeds identically to the standard arm. The private arm is identical to the genomic arm, but the pharmacogenomic regression model is replaced with a differentially-private model.

To simulate realistic dosing increments, we assume any combination of three pills from those available at most pharmacies: 0.5, 1, 2, 2.5, 3, 4, 5, 6, 7, and 7.5 mg. The maximum dose was set to 15 mg/day, with possible dose combinations ranging from 0 to 15 mg in 0.5 mg increments.

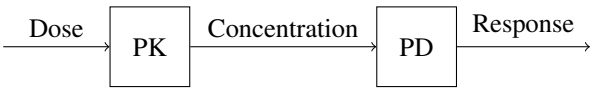

Figure 7: Basic functionality of PK/PD modelling.

5.4 PK/PD Model for INR response to Warfarin

A PK/PD model integrates two distinct pharmacologic models—pharmacokinetic (PK) and pharmacodynamic (PD)—into a single set of mathematical expressions that predict the intensity of a subject's response to drug administration over time. *Pharmacokinetics* is the course of drug absorption, distribution, metabolism, and excretion over time. Mechanistically, the pharmacokinetic component of a PK/PD model predicts the *concentration* of a drug in certain parts of the body over time. *Pharmacodynamics* refers to the effect that a drug has on the body, given its concentration at a particular site. This includes the intensity of its therapeutic and toxic effects, which is the role of the pharmacodynamic component of the PK/PD model. Conceptually, these pieces fit together as shown in Figure 7: the PK model takes a sequences of doses, produces a prediction of drug concentraation, which is given to the PD model. The final output is the predicted PD response to the given sequence of doses, both measures being taken over time. The input/output behavior of the model's components can be described as

the following related functions:

$$\begin{aligned} \mathsf{PKPDModel}(\mathit{genotype}, \mathit{demographics}) &\mapsto F_{\text{inr}} \\ F_{\text{inr}}(\mathit{doses}, \mathit{time}) &\mapsto \mathit{INR} \end{aligned}$$

The function PKPDModel transforms a set of patient characteristics, including the relevant genotype and demographic information, into an INR-response predictor F_{inr}. $F_{\text{inr}}(\mathit{doses}, t)$ transforms a sequence of doses, assumed to have been administered at 24-hour intervals starting at $\mathit{time} = 0$, as well as a time t, and produces a prediction of the patient's INR at time t. The function PKPDModel can be thought of as the routine that initializes the parameters in the PK and PD models, and F_{inr} as the function that composes the initialized models to translate dose schedules into INR measurements. For further details of the PK/PD model, consult Appendix A.

5.5 Calculating Patient Risk

INR levels correspond to the coagulation tendency of blood, and thus to the risk of adverse events. Sorensen *et al.* performed a pooled analysis of the correlation between stroke and bleeding events for patients undergoing warfarin treatment at varying INR levels [41]. We use the probabilities for various events as reported in their analysis. We calculate each simulated patient's risk for stroke, intra-cranial hemorrhage, extra-cranial hemorrhage, and fatality based on the predicted INR levels produced by the PK/PD model. At each 24-hour interval, we calculated INR and the corresponding risk for these events. The sum total risk for each event across the entire trial period is the endpoint we use to compare the arms. We also calculated the mean *time in therapeutic range* (TTR) of patients' INR response for each arm. We define TTR as any INR reading between 1.8–3.2, to maintain consistency with previous studies [3, 14].

The results are presented in Figure 8 in terms of relative risk (defined as the quotient of the patient's risk for a certain outcome when using a particular algorithm versus the fixed dose algorithm). The results are striking: for reasonable privacy budgets ($\varepsilon \leq 5$), private pharmacogenomic dosing results in greater risk for stroke, bleeding, and fatality events as compared to the fixed dose protocol. The increased risk is statistically significant for both private algorithms up to $\varepsilon = 5$ and all types of risk (including reduced TTR), except for private histograms, for which there was no significant increase in bleeding events with $\varepsilon > 1$.

On the positive side, there is evidence that both algorithms may reduce all types of risk at certain privacy levels. Differentially-private histograms performed slightly better, improvements in all types of risk at $\varepsilon \geq 20$. Private linear regression seems to yield lower risk of stroke and fatality and increased TTR at $\varepsilon \geq 20$. However, the difference in bleeding risk for DPLR was not statistically significant at any $\varepsilon \geq 20$. *These results lead us to conclude that there is evidence that differentially-private statistical models may provide effective algorithms for predicting initial warfarin dose, but only at low settings of* $\varepsilon \geq 20$ *that yield little privacy (see Section 4).*

6 Related Work

The tension between privacy and data utility has been explored by several authors. Brickell and Shmatikov [6] found strong evidence for a tradeoff in attribute privacy and predictive performance in common data mining tasks when k-anonymity, ℓ-diversity, and t-closeness are applied before releasing a full dataset. Differential privacy arose partially as a response to Dalenius' desideratum: *anything that can be learned from the database about a specific individual should be learnable without access to the database* [9]. Dwork showed the impossibility of achieving this result in the presence of utility requirements [11], and proposed an alternative goal that proved feasible to achieve in many settings: *the risk to one's privacy should not substantially increase as a result of participating in a statistical database*. Differential privacy formalizes this goal, and constructive research on the topic has subsequently flourised.

Differential privacy is often misunderstood by those who wish to apply it, as pointed out by Dwork and others [13]. Kifer and Machanavajjhala [25] addressed several common misconceptions about the topic, and showed that under certain conditions, it fails to achieve a privacy goal related to Dwork's: *nearly all evidence of an individual's participation should be removed*. Using hypothetical examples from social networking and census data release, they demonstrate that when rows in a database are correlated, or when previous exact statistics for a dataset have been released, this notion of privacy may be violated even when differential privacy is used. Part of our work extends theirs by giving a concrete examples from a realistic application where common misconceptions about differential privacy lead to surprising privacy breaches, i.e., that it will protect genomic attributes from unwanted disclosure. We further extend their analysis by providing a quantitative study of the tradeoff between privacy and utility in the application.

Others have studied the degree to which differential privacy leaks various types of information. Cormode showed that if one is allowed to pose certain queries relating sensitive attributes to quasi-identifiers, it is possible to build a differentially-private Naive Bayes classifier that accurately predicts the sensitive attribute [8]. In contrast, we show that given a model for predicting a

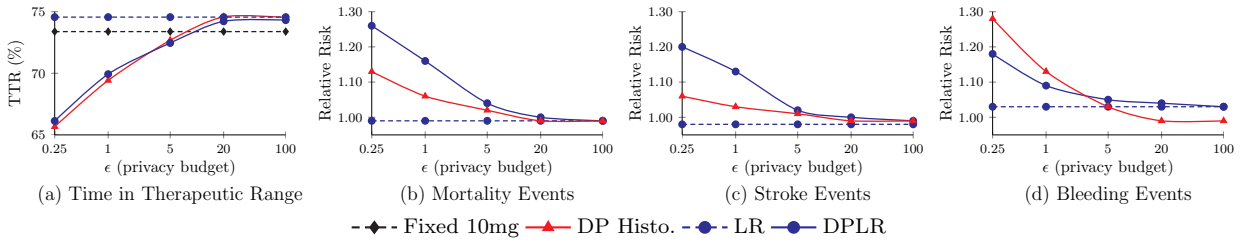

Figure 8: Trial outcomes for fixed dose, non-private linear regression (LR), differentially-private linear regression (DPLR), and private histograms. Horizontal axes represent ε.

certain outcome from a set of inputs (and no control over the queries used to construct the model), it is possible to make accurate predictions in the *reverse* direction: predict one of the inputs given a subset of the other values. Lee and Clifton [30] recognize the problem of setting ε and its relationship to the *relative* nature of differential privacy, and later [31] propose an alternative parametization of differential privacy in terms of the *probability that an individual contributes to the resulting model.* While this may make the privacy guarantee easier for non-specialists to understand, its close relationship to the standard definition suggests that it may not be effective at mitigating the types of disclosures documented in this paper; evaluating its efficacy remains future work, as we are not aware of any existing implementations that support their definition.

The risk of sensitive information disclosure in medical studies has been examined by many. Wang *et al.* [46], Homer *et al.* [20] and Sankararaman *et al.* [39] show that it is possible to recover parts of an individual's genotype given partial genetic information and detailed statistics from a GWAS. They do not evaluate the efficacy of their techniques against private versions of the statistics, and do not consider the problem of inference from a model derived from the statistics. Sweeny showed that a few pieces of identifying information are suitable to identify patients in medical records [42]. Loukides *et al.* [34] show that it is possible to identify a wide range of sensitive patient information from de-identified clinical data presented in a form standard among medical researchers, and later proposed a domain-specific utility-preserving scheme similar to k-anonymity for mitigating these breaches [35]. Dankar and Emam [10] discuss the use of differential privacy in medical applications, pointing out the various tradeoffs between interactive and non-interactive mechanisms and the limitation of utility guarantees in differential privacy, but do not study its use in any specific medical applications.

Komarova *et al.* [28] present an in-depth study of the problem of *partial disclosure*. There is some similarity between the model inversion attacks discussed here and this notion of partial disclosure. One key difference is that in the case of model inversion, an adversary is given the actual function corresponding to a statistical estimator (e.g., a linear model in our case study), whereas Komarova *et al.* consider static estimates from combined public and private sources. In the future we will investigate whether the techniques described by Komarova *et al.* can be used to refine, or provide additional information for, model inversion attacks.

7 Conclusion

We conducted the first end-to-end case study of the use of differential privacy in a medical application, exploring the tradeoff between privacy and utility that occurs when existing differentially-private algorithms are used to guide dosage levels in warfarin therapy. Using a new technique called *model inversion*, we repurpose pharmacogenetic models to infer patient genotype. We showed that models used in warfarin therapy introduce a threat to patients' genomic privacy. When models are produced using state-of-the-art differential privacy mechanisms, genomic privacy is protected for small $\varepsilon(\leq 1)$, but as ε increases towards larger values this protection vanishes.

We evaluated the *utility* of differential privacy mechanisms by simulating clinical trials that use private models in warfarin therapy. This type of evaluation goes beyond what is typical in the literature on differential privacy, where raw statistical accuracy is the most common metric for evaluating utility. We show that differential privacy substantially interferes with the main purpose of these models in personalized medicine: for ε values that protect genomic privacy, which is the central privacy concern in our application, the risk of negative patient outcomes increases beyond acceptable levels.

Our work provides a framework for assessing the tradeoff between privacy and utility for differential privacy mechanisms in a way that is directly meaningful for specific applications. For settings in which certain levels of utility performance *must* be achieved, and this tradeoff cannot be balanced, then alternative means of protecting individual privacy must be employed.

References

[1] Clarification of optimal anticoagulation through genetics. `http://coagstudy.org`.

[2] The pharmacogenomics knowledge base. `http://www.pharmgkb.org`.

[3] J. L. Anderson, B. D. Horne, S. M. Stevens, A. S. Grove, S. Barton, Z. P. Nicholas, S. F. Kahn, H. T. May, K. M. Samuelson, J. B. Muhlestein, J. F. Carlquist, and for the Couma-Gen Investigators. Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. *Circulation*, 116(22):2563–2570, 2007.

[4] P. L. Bonate. Clinical trial simulation in drug development. *Pharmaceutical Research*, 17(3):252–256, 2000.

[5] L. D. Brace. Current status of the international normalized ratio. *Lab Medicine*, 32(7):390–392, 2001.

[6] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *KDD*, 2008.

[7] J. Carlquist, B. Horne, J. Muhlestein, D. Lapp, B. Whiting, M. Kolek, J. Clarke, B. James, and J. Anderson. Genotypes of the Cytochrome P450 Isoform, CYP2C9, and the Vitamin K Epoxide Reductase Complex Subunit 1 conjointly determine stable warfarin dose: a prospective study. *Journal of Thrombosis and Thrombolysis*, 22(3), 2006.

[8] G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *KDD*, 2011.

[9] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1, 1977.

[10] F. K. Dankar and K. El Emam. The application of differential privacy to health data. In *ICDT*, 2012.

[11] C. Dwork. Differential privacy. In *ICALP*. Springer, 2006.

[12] C. Dwork. The promise of differential privacy: A tutorial on algorithmic techniques. In *FOCS*, 2011.

[13] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Differential privacy: A primer for the perplexed. In *Joint UNECE/Eurostat work session on statistical data confidentiality*, 2011.

[14] V. A. Fusaro, P. Patil, C.-L. Chi, C. F. Contant, and P. J. Tonellato. A systems approach to designing effective clinical trials using simulations. *Circulation*, 127(4):517–526, 2013.

[15] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, 2008.

[16] A. K. Hamberg, Dahl, M. L., M. Barban, M. G. Srordo, M. Wadelius, V. Pengo, R. Padrini, and E. Jonsson. A PK-PD model for predicting the impact of age, CYP2C9, and VKORC1 genotype on individualization of warfarin therapy. *Clinical Pharmacology Theory*, 81(4):529–538, 2007.

[17] D. Hand and R. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

[18] N. Holford, S. C. Ma, and B. A. Ploeger. Clinical trial simulation: A review. *Clinical Pharmacology Theory*, 88(2):166–182.

[19] N. H. G. Holford, H. C. Kimko, J. P. R. Monteleone, and C. C. Peck. Simulation of clinical trials. *Annual Review of Pharmacology and Toxicology*, 40(1):209–234, 2000.

[20] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8), 08 2008.

[21] International Warfarin Pharmacogenetic Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[22] E. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9), Sept 1982.

[23] F. Kamali and H. Wynne. Pharmacogenetics of warfarin. *Annual Review of Medicine*, 61(1):63–75, 2010.

[24] S. P. Kasiviswanathan, M. Rudelson, and A. Smith. The power of linear reconstruction attacks. In *SODA*, 2013.

[25] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, 2011.

[26] M. J. Kim, S. M. Huang, U. A. Meyer, A. Rahman, and L. J. Lesko. A regulatory science perspective on warfarin therapy: a pharmacogenetic opportunity. *J Clin Pharmacol*, 49:138–146, Feb 2009.

[27] S. E. Kimmel, B. French, S. E. Kasner, J. A. Johnson, J. L. Anderson, B. F. Gage, Y. D. Rosenberg, C. S. Eby, R. A. Madigan, R. B. McBane, S. Z. Abdel-Rahman, S. M. Stevens, S. Yale, E. R. Mohler, M. C. Fang, V. Shah, R. B. Horenstein, N. A. Limdi, J. A. Muldowney, J. Gujral, P. Delafontaine, R. J. Desnick, T. L. Ortel, H. H. Billett, R. C. Pendleton, N. L. Geller, J. L. Halperin, S. Z. Goldhaber, M. D. Caldwell, R. M. Califf, and J. H. Ellenberg. A pharmacogenetic versus a clinical algorithm for warfarin dosing. *New England Journal of Medicine*, 369(24):2283–2293, 2013. PMID: 24251361.

[28] T. Komarova, D. Nekipelov, and E. Yakovlev. *Estimation of Treatment Effects from Combined Data: Identification versus Data Security*. NBER volume Economics of Digitization: An Agenda, To appear.

[29] M. J. Kovacs, M. Rodger, D. R. Anderson, B. Morrow, G. Kells, J. Kovacs, E. Boyle, and P. S. Wells. Comparison of 10-mg and 5-mg warfarin initiation nomograms together with low-molecular-weight heparin for outpatient treatment of acute venous thromboembolism. *Annals of Internal Medicine*, 138(9):714–719, 2003.

[30] J. Lee and C. Clifton. How much is enough? Choosing ε for differential privacy. In *ISC*, 2011.

[31] J. Lee and C. Clifton. Differential identifiability. In *KDD*, 2012.

[32] J. Lei. Differentially private m-estimators. In *NIPS*, 2011.

[33] Y. Lindell and E. Omri. A practical application of differential privacy to personalized online advertising. *IACR Cryptology ePrint Archive*, 2011.

[34] G. Loukides, J. C. Denny, and B. Malin. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, 17(3):322–327, 2010.

[35] G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(17):7898–7903, Apr. 2010.

[36] A. Narayanan and V. Shmatikov. Robust deanonymization of large sparse datasets. In *Oakland*, 2008.

[37] A. Narayanan and V. Shmatikov. Myths and fallacies of *Personally Identifiable Information*. *Commun. ACM*, 53(6), June 2010.

[38] J. Reed, A. J. Aviv, D. Wagner, A. Haeberlen, B. C. Pierce, and J. M. Smith. Differential privacy for collaborative security. In *Proceedings of the Third European Workshop on System Security*, EUROSEC, 2010.

[39] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9):965–967, 2009.

[40] E. A. Sconce, T. I. Khan, H. A. Wynne, P. Avery, L. Monkhouse, B. P. King, P. Wood, P. Kesteven, A. K. Daly, and F. Kamali. The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood*, 106(7):2329–2333, 2005.

[41] S. V. Sorensen, S. Dewilde, D. E. Singer, S. Z. Goldhaber, B. U. Monz, and J. M. Plumb. Cost-effectiveness of warfarin: Trial versus real-world stroke prevention in atrial fibrillation. *American Heart Journal*, 157(6):1064 – 1073, 2009.

[42] L. Sweeney. Simple demographics often identify people uniquely. 2000.

[43] F. Takeuchi, R. McGinnis, S. Bourgeois, C. Barnes, N. Eriksson, N. Soranzo, P. Whittaker, V. Ranganath, V. Kumanduri, W. McLaren, L. Holm, J. Lindh, A. Rane, M. Wadelius, and P. Deloukas. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet*, 5(3), 03 2009.

[44] S. Vinterbo. Differentially private projected histograms: Construction and use for prediction. In *ECML-PKDD*, 2012.

[45] D. Vu and A. Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *ICDM Workshops*, 2009.

[46] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: information leaks in genome wide association studies. In *CCS*, 2009.

[47] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: regression analysis under differential privacy. In *VLDB*, 2012.

A PK/PD Model Details

We adopted a previously-developed PK/PD INR model to predict each patient's INR response to previous dosing choices [16]. The PK component of the model is a *two-compartment* model with *first-order absorption*. A two-compartment model assumes an abstract representation of the body as two discrete sections: the first being a *central* compartment into which a drug is administered and a *peripheral* compartment into which the drug eventually distributes. The central compartment (assumed to have volume V_1) represents tissues that equilibrate rapidly with blood (e.g., liver, kidney, *etc.*), and the peripheral (volume V_2) those that equilibrate slowly (e.g., muscle, fat, *etc.*). Three *rate constants* govern transfer between the compartments and elimination: k_{12}, k_{21}, for the central-peripheral and peripheral-central transfer, and k_{el} for elimination from the body, respectively. V_1, V_2, k_{12}, and k_{21} are related by the following equality: $V_1 k_{12} = V_2 k_{21}$. The *absorption rate* k_a governs the rate at which the drug enters the central compartment. In the model used in our simulation, each of these parameters is represented by a random variable whose distribution has been fit to observed population measurements of Warfarin absorption, distribution, metabolism, and elimination [16]. The elimination-rate constant k_{el} is parameterized by the patient's CYP2C9 genotype.

Given a set of PK parameters, the Warfarin concentration in the central compartment over time is calculated using standard two-compartment PK equations for oral dosing. Concentration in two-compartment pharmacokinetics diminishes in two distinct phases with differing rates: the α ("distribution") phase, and β ("elimination") phase. The expression for concentration C over time assuming doses $D_1, \ldots, D_n$ administered at times $t_{D_1}, \ldots, t_{D_n}$ has another term corresponding to the effect of oral absorption:

$$C(t) = \sum_{i=1}^{n} D_i (A e^{-\alpha t_i} + B e^{-\beta t_i} - (A+B) e^{-k_a t_i})$$

with $t_i = t - t_{D_i}$ and α, β satisfying $\alpha\beta = k_{21} k_{el}$, $\alpha + \beta = k_{el} + k_{12} + k_{21}$, and

$$A = \frac{k_a}{V_1} \frac{k_{21} - \alpha}{(k_a - \alpha)(\beta - \alpha)} \quad B = \frac{k_a}{V_1} \frac{k_{21} - \beta}{(k_a - \beta)(\alpha - \beta)}$$

Our model contains an error term with a zero-centered log-normal distribution whose variance depends on whether or not steady-state dosing has occurred; the term is given in the appendix of Hamberg *et al.* [16].

PD Model The PD model used in our simulations is an *inhibitory sigmiod-$E_{\max}$ model*. Recall that the purpose of the PD model is to describe the physiological response E, in this case INR, to Warfarin concentration at a particular time. $E_{\max}$ represents the maximal response, i.e., the maximal inhibition of coagulation, and E_{50} the concentration of Warfarin producing half-maximal inhibition. $E_{\max}$ is fixed to 1, and E_{50} is a patient-specific random variable that is a function of the patient's VKORC1 genotype. A sigmoidocity factor γ is used to model the fact that the concentration-effect response of Warfarin corresponds to a sigmoid curve at lower concentrations. The basic formula for calculating E at time t from concentration is: $1 - (E_{\max} C(t)^\gamma)/(E_{50}^\gamma + C(t)^\gamma)$. However, Warfarin exhibits a delay between exposure and anticoagulation response. To characterize this feature, Hamberg *et al.* showed that extending the basic $E_{\max}$ model with a *transit compartment model* with two parallel chains is adequate [16], as shown in Figure 9. The delay between exposure and concentration is modeled by assuming that the drug travels along two parallel *compartment chains* of differing lengths and turnover rates. The transit rate between compartments on the two chains is given by two constants k_{tr1} and k_{tr2}. The first chain consists of six compartments, and the second a single compartment. The first transit constant is a random zero-centered log-normal variable, whereas empirical data did not reilably support variance in the second [16]. The amount in a given compartment i, A_i, at time t is described by a system of coupled ordinary differential equations:

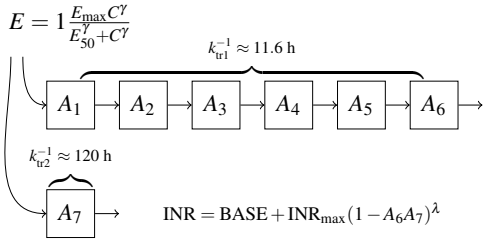

Figure 9: Overview of transit-compartment PD model [16].

$$\frac{dA_1}{dt} = k_{tr1} \left(1 - \frac{E_{\max} C(t)^\gamma}{E_{50}^\gamma + C(t)^\gamma}\right) - k_{tr1} A_1$$

$$\frac{dA_n}{dt} = k_{tr1}(A_{n-1} - A_n), n = 2,3,4,5,6$$

$$\frac{dA_7}{dt} = k_{tr2} \left(1 - \frac{E_{\max} C(t)^\gamma}{E_{50}^\gamma + C(t)^\gamma}\right) - k_{tr2} A_7$$

The final expression for INR at time t is given by solving for A_7 and A_7 starting from initial conditions $A_i = 1$, and calculating the expression: $\log(\text{INR}) = \log(\text{Base} + \text{INR}_{\max}(1 - A_6 A_7)^\lambda) + \varepsilon_{\text{INR}}$. In this expression, Base is the patient's baseline INR, $\text{INR}_{\max}$ is the maximal INR (assumed to be 20 [16]), λ is a scaling factor derived from empirical data [16], and ε_{INR} is a zero-centerd, symmetrically-distributed random variable with variance determined from empirical data [16].