

Interpretable Subgroup Discovery in Treatment Effect Estimation with Application to Opioid Prescribing Guidelines

Chirag Nagpal,^{1,2} Dennis Wei,¹ Bhanukiran Vinzamuri,¹ Monica Shekhar,³

Sara E. Berger,¹ Subhro Das,¹ and Kush R. Varshney¹

¹IBM Research

²Carnegie Mellon University

³IBM Global Business Services

ABSTRACT

The dearth of prescribing guidelines for physicians is one key driver of the current opioid epidemic in the United States. In this work, we analyze medical and pharmaceutical claims data to draw insights on characteristics of patients who are more prone to adverse outcomes after an initial synthetic opioid prescription. Toward this end, we propose a generative model that allows discovery from observational data of subgroups that demonstrate an enhanced or diminished causal effect due to treatment. Our approach models these sub-populations as a mixture distribution, using sparsity to enhance interpretability, while jointly learning nonlinear predictors of the potential outcomes to better adjust for confounding. The approach leads to human-interpretable insights on discovered subgroups, improving the practical utility for decision support.

CCS CONCEPTS

• Applied computing → Health informatics; • Mathematics of computing → Causal networks; Bayesian networks; Probabilistic representations.

KEYWORDS

causal treatment effect, heterogeneous treatment effect, Bayesian networks, decision support

ACM Reference Format:

Chirag Nagpal,^{1,2} Dennis Wei,¹ Bhanukiran Vinzamuri,¹ Monica Shekhar,³ Sara E. Berger,¹ Subhro Das,¹ and Kush R. Varshney¹, ¹IBM Research, ²Carnegie Mellon University, ³IBM Global Business Services . 2020. Interpretable Subgroup Discovery in Treatment Effect Estimation with Application to Opioid Prescribing Guidelines. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20)*, April 2–4, 2020, Toronto, ON, Canada. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3368555.3384456>

1 INTRODUCTION

The United States is in the midst of an opioid addiction epidemic. According to estimates by the Centers for Disease Control and Prevention (CDC), 42,000 people died from opioid overdoses in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7046-2/20/04.

<https://doi.org/10.1145/3368555.3384456>

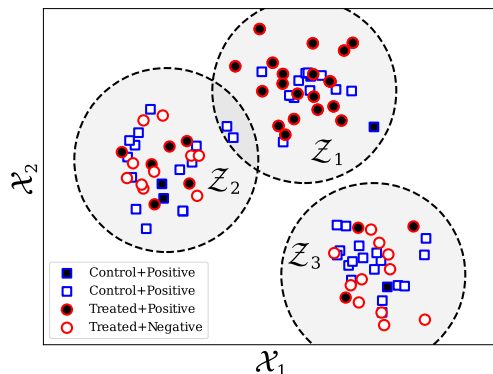


Figure 1: The heterogeneous effect subgroup discovery problem. Almost all instances receiving treatment in Z_1 have a positive outcome, while very few in Z_3 do. We are interested in recovering such latent subgroups.

2016 and 49,000 in 2017. Overdose mortalities specifically from *synthetic* opioids, such as Fentanyl, have increased exponentially since 1999.¹ A major cause of this epidemic is overprescription of opioids (for legitimate pain management) by physicians who lack proper prescribing guidelines [5, 29].

One actionable insight for prescribers is characteristics of patients for whom treatment with synthetic opioids, as opposed to natural or semi-synthetic opioids, causes a greater risk of adverse outcomes such as long-term use and addiction than for the general population. Toward this end, we study causal treatment effect estimation from observational data under heterogeneity, i.e. the phenomenon of different individuals having different responses to the same treatment. In particular, we focus on the discovery of subgroups of patients (really portions of a feature space) that have enhanced or diminished treatment effects. We aim for the discovered subgroups to be human-interpretable so that the results can be directly used in prescribing guidelines.

We make use of the MarketScan database of medical claims and pharmaceutical claims. Claims are a form of administrative data that are commonly repurposed for medical studies because they capture the diagnoses, procedures, and prescriptions of patients longitudinally. Inference of causal relationships using observational data is challenging since counterfactual outcomes are not observed and the treated and untreated populations may have underlying

¹<https://www.cdc.gov/drugoverdose/data/analysis.html>

differences that affect the outcome. However, modern machine learning techniques provide an avenue to overcome the challenges.

To identify subgroups with different treatment effects, we hypothesize that a latent variable determines the treatment effect of each individual. Moreover, individuals with similar characteristics belong to the same latent subgroup, resulting in similar responses to treatment across the subgroup. Figure 1 is an abstract representation of such a phenomenon.

We propose a Bayesian network to model these subgroups, specifically as a mixture model, along with their corresponding treatment effects. Sparsity is induced in the learned mixture component parameters to improve interpretability. Our approach, which we name the heterogeneous effect mixture model (HEMM), is similar in spirit to causal rule sets for identifying subgroups with enhanced treatment effect [48] but does not require hard partitions or assignments. Moreover, we incorporate nonlinear as well as linear outcome models, which increases the expressiveness of the model to better adjust for confounding without sacrificing the interpretability of the subgroup definitions. We thus benefit from both the interpretability of sparse mixture models and the representation learning capability of neural networks. In contrast, recent works [1, 28, 43] that use neural networks or nonparametric methods to estimate heterogeneous treatment effects do not identify subgroups of individuals with similar responses. While our motivating application is opioid use, the proposed approach applies to any problem domain requiring the discovery of subgroups with heterogeneous responses to actions. In this spirit, we also validate our method on synthetic data and the Infant Health and Development Program (IHDP) dataset in terms of its heterogeneous effect estimation and subgroup identification performance.

With respect to opioids, we provide domain expert interpretation of the enhanced treatment effect subgroup discovered using MarketScan data, i.e. patients at higher risk of adverse outcomes after an initial synthetic opioid prescription. Some characteristics of this subgroup are well-known and/or reflected in CDC opioid prescribing guidelines [9]: chronic pain conditions, psychological comorbidities, heart disease and obesity. The presence of minor injuries and dental/oral conditions in the subgroup can be explained by the common practice of prescribing opioids for post-surgical or intense acute pain. Lastly, some discovered conditions are unexpected, such as skin infections, abscesses, and reproductive disorders.

Overall, our contributions can be summarized as follows:

- i) We propose the HEMM for discovering subgroups with enhanced and diminished treatment effects in a potential outcomes causal inference framework, using sparsity to enhance interpretability.
- ii) We extend the HEMM's outcome model to include neural networks to better adjust for confounding and develop a joint inference procedure for the overall graphical model and the neural networks.
- iii) We demonstrate strong performance in estimating heterogeneous effects and identifying subgroups compared to existing approaches.
- iv) We apply the methodology to a large-scale medical claims dataset and discover actionable patient subgroups at enhanced risk of adverse outcomes with synthetic opioids.

2 RELATED WORK

There is a rich literature of data-oriented research on understanding the patterns and risks of opioid prescribing and addiction in the fields of medicine, medical informatics, and machine learning; some is specifically intended to inform prescribing guidelines. Kim et al. [22] conduct a randomized controlled trial and Neill and Herlands [32] analyze spatiotemporal overdose event data, but a large part of the literature works with medical claims data [3, 11, 20, 24, 35, 50] and similar administrative data [6] as we do. However, unlike us, none of these works focus on heterogeneous treatment effects.

The identification of subgroups with heterogeneous or enhanced treatment effects has been addressed in the statistics literature by building separate factual and counterfactual outcome models and then regressing the difference of the two using another method, e.g. a decision tree [46]. This final model can then be deployed to identify subgroups. Within this category of approaches, Lipkovich et al. [27] propose the subgroup identification based on differential effect search (SIDES) algorithm, Dusseldorp and Van Mechelen [10] propose the qualitative interaction trees (QUINT) algorithm, and Foster et al. [12] propose the virtual twins (VT) method. We consider empirical comparisons to these algorithms in the sequel.

Wang and Rudin [48] propose causal rule sets for discovering subgroups with enhanced treatment effect. This is the closest to and an inspiration for our work. That work seeks to learn discrete human-interpretable rules predictive of enhanced treatment effect and involves optimization by Monte Carlo methods. We consider instead a mixture of experts approach with soft assignment to groups that retains most of the interpretability but allows greater expressiveness and can be optimized via gradient methods. Our outcome model (6), (7) also differs from that of [48]. Most importantly, we allow nonlinearity in the form of neural networks whereas [48] considers only linear models. Our model also has a single term representing the main effect of treatment whereas [48] has three such terms: a population average, a subgroup term that is always active, and a subgroup term that is only active under treatment.

Recent papers have proposed estimating heterogeneous/individual treatment effects using neural networks [28, 43] or a Bayesian nonparametric method involving Gaussian processes [1]. These methods rely on constructing distributional representations of the factual and counterfactual outcomes that are similar in a statistical sense. While these methods perform well on estimating heterogeneous effects, they do not identify subgroups of individuals with similar treatment effects and characteristics and are thus less interpretable. This makes the application of such methods to inform policy decisions more difficult.

3 HETEROGENEOUS EFFECT MIXTURE MODEL

In this section, we propose a generative mixture model for heterogeneous treatment effects. One way to model heterogeneity, and the one in our proposal, is as a finite mixture of components with a different treatment effect model in each component (some enhanced and some diminished). For tractability, we keep the form of the mixtures to be the simplest possible: Gaussian-distributed for continuous covariates and Bernoulli-distributed for discrete covariates. We encode a preference for components to involve few covariates

through a Laplace prior or a group $\ell_{1,2}$ prior on the means of the covariates, described in Section 3.3. Section 3.4 presents a model for the outcomes as they depend on treatment, covariates, and mixture membership, including nonlinear dependence on the covariates.

3.1 Preliminaries

We adopt the Neyman-Rubin potential outcomes framework [40] for causal inference. Define random variables $\mathbf{X} \in \mathbb{R}^d$ representing covariates and $T \in \{0, 1\}$ as the treatment indicator. The subset of continuous-valued covariates are denoted \mathbf{X}_{cont} and the discrete covariates (binary-valued or binarized) are denoted \mathbf{X}_{disc} . We will sometimes refer to $T = 1$ as ‘the treatment’ and $T = 0$ as ‘the control.’ Corresponding to the levels of treatment are two potential outcomes $Y(0)$ and $Y(1)$, which are the outcomes under $T = 0$ and $T = 1$ respectively. These outcomes can be discrete- or continuous-valued. We are given an observational dataset of samples $\mathcal{D} = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^N$ in which only one of the outcomes is observed for each individual: if $t_i = 0$ then $y_i = y(0)_i$, and if $t_i = 1$ then $y_i = y(1)_i$.

Our interest lies in estimating the conditional average treatment effect (CATE) conditioned on \mathbf{X} , defined as

$$\tau(\mathbf{X}) = \mathbb{E}[Y(1) - Y(0) | \mathbf{X}].$$

In this work, the dependence on \mathbf{X} is mediated primarily through subgroup membership, i.e. members of the same subgroup have similar treatment effects. We make the standard assumptions that allow CATE to be identifiable from observational data, namely exchangeability conditioned on the available covariates, $T \perp (Y(0), Y(1)) | \mathbf{X}$, positivity of the treatment propensity, $0 < p(T = 1 | \mathbf{x}) < 1$ for all \mathbf{x} , and no dependence between individuals i.e. the stable unit treatment value assumption (SUTVA) [17, 41]. The first two assumptions are collectively known as strong ignorability (SITA).

For the mixture model proposed in this paper, we additionally define the latent random variable $Z \in \mathcal{Z} = \{1, \dots, K\}$ to indicate mixture membership. Both the distribution of covariates and the treatment effect are dependent on Z as described next.

3.2 Generative Model

The generative model is presented in Figure 2 in plate notation. We first give an overview of the distributions and then go into more detail regarding \mathbf{X} and Y in Sections 3.3 and 3.4.

- (1) We draw a sample z_i independently for each individual i that determines the latent group membership. The prior distribution for Z is uniform over the K groups,

$$Z \sim \text{Uniform}(K). \quad (1)$$

- (2) Conditioned on the latent group assignment $z_i = k$,
 - (a) The $\mathbf{x}_{\text{cont}, i}$ are drawn i.i.d. from a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance Σ_k :

$$\mathbf{X}_{\text{cont}} | z_i = k \sim \text{Normal}(\boldsymbol{\mu}_k, \Sigma_k). \quad (2)$$

In this paper, we constrain the off-diagonal elements of Σ_k to be 0 to reduce the number of parameters, although non-diagonal covariances can be easily accommodated.

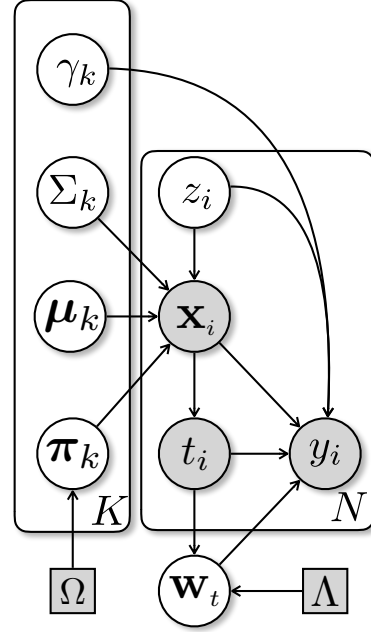


Figure 2: The proposed heterogeneous effect mixture model (HEMM) in plate notation. For each instance i , (\mathbf{x}_i, t_i, y_i) are the observed variables and z_i is a latent variable that determines membership in one of the K mixture components. Each component has an associated coefficient γ_k that determines the main treatment effect.

- (b) The $\mathbf{x}_{\text{disc}, i}$ are drawn i.i.d. from a multivariate Bernoulli distribution with mean $\boldsymbol{\pi}_k$:

$$\mathbf{X}_{\text{disc}} | z_i = k \sim \text{Bernoulli}(\boldsymbol{\pi}_k). \quad (3)$$

We enforce sparsity in $\boldsymbol{\pi}_k$ in order to improve interpretability. We describe this in detail in Section 3.3.

- (3) Conditioned on the covariates \mathbf{x}_i , the treatment assignment t_i is drawn from a Bernoulli distribution whose mean is a function of \mathbf{x}_i :

$$T | \mathbf{x} \sim \text{Bernoulli}(p(\mathbf{x})).$$

This corresponds to a model for treatment *propensity*. Note from Figure 2 that the generative model assumes that T is conditionally independent of Z given \mathbf{X} . Under this assumption, it will be seen in Section 4 that inference for the propensity model can be done independently from the other components of the generative model.

- (4) Finally, an outcome sample y_i is drawn from a distribution whose mean μ_y is a function of the covariates \mathbf{x}_i , treatment assignment t_i , and latent group assignment z_i . If Y is binary-valued, the distribution is Bernoulli,

$$Y | \mathbf{x}, t, z \sim \text{Bernoulli}(\mu_y(\mathbf{x}, t, z)),$$

whereas if Y is continuous, the distribution is Gaussian,

$$Y | \mathbf{x}, t, z \sim \text{Normal}(\mu_y(\mathbf{x}, t, z), \sigma_y^2),$$

where σ_y^2 is the variance. The outcome model is discussed further in Section 3.4.

Note we are interested in estimating the causal quantity,

$$\mathbb{E}[Y(t)|X] = \mathbb{E}[Y|\text{do}(T = t), X] = p(Y|\text{do}(T = t), X).$$

Here, the first equality is from definition of interventional quantities and the second equality holds due to Y being binary.

THEOREM 1 (IDENTIFIABILITY). *Under the Directed Acyclic Graph in Figure. 2,*

$$p(Y|\text{do}(T = t), X) = \int_Z p(Y|X, Z, T = t)p(Z|X).$$

Theorem 1 confirms that we can estimate the CATE from the observational quantities introduced above. The proof is deferred to the Appendix A.

3.3 Sparse Mixture Components for Interpretability

Without further measures, the mixture component means μ_k and π_k learned from data may be dense, making them difficult for a domain expert to interpret. We hypothesize that a large number of learned mean parameters may have small values and that promoting sparsity through appropriate prior distributions can overcome this problem. To this end, we experiment with two different sparsity-promoting priors on the means π_k of discrete covariates. The same priors can be placed on the continuous covariate means μ_k but we do not find this necessary in the present work.

- (1) **Laplace (ℓ_1) Prior:** We assume that the means π_{jk} follow zero-mean Laplace distributions and are independent across mixture components and covariates. The negative log-likelihood is therefore proportional to the ℓ_1 norm

$$\Omega(\pi) = \sum_{j \in \text{disc}} \sum_{k=1}^K |\pi_{jk}|, \quad (4)$$

where the summation over j is restricted to the discrete covariates.

- (2) **Group $\ell_{1,2}$ Prior:** It may further be the case that some covariates are non-informative of group membership, in which case the means π_{jk} should be zero across all groups k and follow the group $\ell_{1,2}$ distribution [30], similar to the group lasso [49]. The corresponding negative log-likelihood is

$$\Omega(\pi) = \sum_{j \in \text{disc}} \sqrt{\sum_{k=1}^K |\pi_{jk}|^2}. \quad (5)$$

3.4 Treatment Outcome Model

We model the enhanced or diminished treatment effect in a subgroup through the following relationships. In the case where Y is binary, its mean is equal to the probability of $Y = 1$. We define the latter using the logistic sigmoid function g to be

$$p(Y = 1 | \mathbf{x}, t, Z = k; \mathbf{w}_t, \gamma_k) = g(f(\mathbf{x}; \mathbf{w}_t) + \gamma_k t), \quad (6)$$

where $f(\mathbf{x}; \mathbf{w}_t)$ is a function of \mathbf{x} parametrized by \mathbf{w}_t , $t = 0, 1$. The term $\gamma_k t$ represents the main effect due to treatment and the coefficient γ_k , i.e. the size of the effect, depends on the group membership $Z = k$. The parameters \mathbf{w}_t are allowed to be different for $t = 0$ and

$t = 1$ to better account for differing covariate distributions $p(\mathbf{x} | t)$ between the two treatment groups, a.k.a. selection bias. In the case of continuous Y , we replace g with the identity function as follows:

$$\mathbb{E}[Y = 1 | \mathbf{x}, t, Z = k; \mathbf{w}_t, \gamma_k] = f(\mathbf{x}; \mathbf{w}_t) + \gamma_k t. \quad (7)$$

The simplest choice for function $f(\cdot)$ is linear, i.e., two linear functions $\mathbf{w}_0^\top \mathbf{x}$ and $\mathbf{w}_1^\top \mathbf{x}$. In practice, however, the outcome may have a highly nonlinear dependence on the covariates. To accommodate nonlinear covariate interactions and thus better adjust for confounding, we also allow f to be a nonlinear function. In this paper, we experiment with one- and two-hidden-layer feedforward neural networks with ReLU activations. Outcomes under $t = 0$ and $t = 1$ are produced by two different heads of the network, following [21, 28, 43]. Even in the nonlinear case, the assignment of an individual to a subgroup is still described by a mixture model and directly interpretable in terms of the original feature representation, thus preserving interpretability of the discovered subgroups.

It is possible to regularize the outcome models (6), (7) with ℓ_2 or ℓ_1 regularization $\Lambda(\mathbf{w}_t)$, which is equivalent to adding a normal or Laplace prior on the parameter \mathbf{w}_t . In this work however, we use weight decay instead as discussed in Section 4.1.

4 INFERENCE

We would like to fit our proposed model to a given observational dataset \mathcal{D} . Denote by $\Theta = (\{\mu_k, \Sigma_k, \pi_k, \gamma_k\}_{k=1}^K, \mathbf{w})$ the set of all parameters of the model.

We have considered two approaches: maximizing the joint likelihood $p(\mathbf{x}_i, t_i, y_i; \Theta)$, and maximizing the conditional likelihood $p(y_i | \mathbf{x}_i, t_i; \Theta)$. The joint and conditional likelihoods can be related as follows:

$$\sum_{i=1}^N \ln p(\mathbf{x}_i, t_i, y_i) = \sum_{i=1}^N [\ln p(\mathbf{x}_i) + \ln p(t_i | \mathbf{x}_i) + \ln p(y_i | \mathbf{x}_i, t_i)]. \quad (8)$$

The conditional likelihood can be further expanded as

$$\ln p(y_i | \mathbf{x}_i, t_i) = \ln \left(\sum_{k=1}^K p(z_i = k | \mathbf{x}_i) p(y_i | \mathbf{x}_i, t_i, z_i = k) \right), \quad (9)$$

where we have used the conditional independence of Z and T given \mathbf{X} in the first factor on the right-hand side. The resulting first factor $p(z_i = k | \mathbf{x}_i)$ as well as the term $p(\mathbf{x}_i)$ in (8) depend only on the mixture model (1)–(3), to wit $p(\mathbf{x}_i) = \sum_{k=1}^K p(z_i = k) p(\mathbf{x}_i | z_i = k)$ and $p(z_i = k | \mathbf{x}_i) = p(z_i = k) p(\mathbf{x}_i | z_i = k) / p(\mathbf{x}_i)$. The second factor on the right-hand side of (9) depends only on the outcome model (6), (7). The remaining term $p(t_i | \mathbf{x}_i)$ in (8) depends on the propensity model. Since this is the only place where the propensity model appears, its inference is separable from the remainder of the problem, as claimed in Section 3.2. We do not discuss propensity modeling further as it is not the focus of this work.

Although maximizing the joint likelihood (8) results in some closed-form expressions and accordingly easier inference of parameters, we have observed in practice that maximizing the conditional likelihood (9) has superior performance in estimating the potential outcomes $Y(t)$ and treatment effects. Therefore, we pursue this discriminative approach in this work. We do however include the

sparsity-inducing prior on the parameters π_k discussed in Section 3.3. The full objective function is therefore

$$\sum_{i=1}^N \ln p(y_i | \mathbf{x}_i, t_i; \Theta) - \lambda \Omega(\pi), \quad (10)$$

where λ controls the strength of the prior.

4.1 Evidence Lower Bound (ELBO) Optimization

Instead of optimizing the conditional log-likelihood in (10) directly using a gradient method, we choose to lower bound the likelihood with a variational approximation, more commonly known as the Evidence Lower Bound (ELBO) [2]. For any variational distribution $q(Z)$ over the latent variable Z , we have

$$\begin{aligned} \ln p(y_i | \mathbf{x}_i, t_i; \Theta) &= \ln \sum_{k=1}^K p(y_i, z_i = k | \mathbf{x}_i, t_i; \Theta) \\ &= \ln \left(\mathbb{E}_q \left[\frac{p(y_i, z_i | \mathbf{x}_i, t_i; \Theta)}{q(Z)} \right] \right) \\ &\geq \mathbb{E}_q \left[\ln \frac{p(y_i, z_i | \mathbf{x}_i, t_i; \Theta)}{q(Z)} \right] \end{aligned} \quad (11)$$

using Jensen's inequality. Now, replacing $q(Z)$ with $p(z_i | \mathbf{x}_i; \Theta)$ and using (9) (and $Z \perp\!\!\!\perp T | \mathbf{X}$ from Figure 2), we obtain

$$\text{ELBO}(y_i, \mathbf{x}_i, t_i; \Theta) = \sum_{k=1}^K p(z_i = k | \mathbf{x}_i; \Theta) \ln p(y_i | \mathbf{x}_i, t_i, z_i = k; \Theta). \quad (12)$$

We hence substitute (12) in place of $\ln p(y_i | \mathbf{x}_i, t_i; \Theta)$ in (10) and proceed to maximize the objective function using the Adam gradient method [23], a variant of stochastic gradient descent that is a popular choice for non-convex functions like neural networks. The same method is used for both linear and nonlinear f in (6), (7). As noted above, the first factor $p(z_i = k | \mathbf{x}_i; \Theta)$ in (12) depends only on the mixture model parameters in (1)–(3) while the second factor depends only on the outcome model parameters in (6), (7). We enable “weight decay” [26] on the parameters \mathbf{w}_t as a form of regularization. For tractability, we compute the ELBO only over a fixed-size mini-batch of the data before each parameter update. Additional details on the algorithm and parameter initialization can be found in Appendices C and D in the supplement.

We also considered an expectation-maximization (EM) algorithm to maximize (10) as an alternative to ELBO. Our experience however was that ELBO provided better fit in terms of Log-Likelihood and heterogeneous effect estimates in terms of the metric reported in Section 6.2. A full description of the EM method and a comparison to the ELBO optimization is deferred to the Appendix.

5 DATASET DESCRIPTIONS

We demonstrate the performance of HEMM on a synthetic dataset, the semi-synthetic Infant Health and Development Program (IHDP) dataset, and a real-world dataset on opioids. These datasets are described further below.

SYNTHETIC: We take $\mathbf{X} = (X_0, X_1) \in \mathbb{R}^2$ and sample it from a uniform distribution over $\mathcal{X} = [0, 1]^2$. In order to simulate the selection

Total Covariate Dimension	1226	3 Continuous, 1223 Binary
ICD-9 Diagnostic Codes	1013	Binary
CPT Procedure Codes	171	Binary
Hand-Crafted Comorbidities	41	Binary
Daily Morphine Equivalent, Total Number of Visits, Age	3	Continuous

	Addicted (Y=1)	Not-Addicted (Y=0)	Total
Treated (T=1)	2060	19983	22043
Control (T=0)	7269	176156	183425
Total	9329	196139	205468

Table 1: OPIOID Dataset Statistics

bias inherent in observational studies, the treatment variable depends on \mathbf{X} as $T \sim \text{Bernoulli}(0.4)$ for $x_0 < 0.5$ and $\text{Bernoulli}(0.6)$ for $x_0 > 0.5$. The potential outcomes $Y(0)$ and $Y(1)$ are also Bernoulli with means given by the functions of \mathbf{X} shown in Figure 3a. The figure shows that $p(Y(1) = 1 | \mathbf{X}) > p(Y(0) = 1 | \mathbf{X})$, i.e. treatment increases the probability of positive outcome. Note that under the conditional exchangeability assumption we have $p(Y(t) = 1 | \mathbf{X}) = p(Y = 1 | T = t, \mathbf{X})$. We model the effect of the confounders \mathbf{X} by assigning higher probability to the upper triangular region of \mathcal{X} . This together with the distribution of T imply that individuals who are more likely to have positive outcome regardless of treatment (upper triangle) are also more likely to receive treatment (right half-square). Lastly, we model the enhanced treatment effect group as a circular region $\mathcal{S} = \{x : \|x - c\|_2 < r\}$, where $p(Y(1) = 1 | \mathcal{S}) > p(Y(1) = 1 | \mathcal{X} \setminus \mathcal{S})$. We set $c = (\frac{1}{2}, \frac{1}{2})$ and $r = \frac{1}{4}$. A total of 1000 samples (\mathbf{x}_i, t_i, y_i) were generated as described above.

IHDP (SEMI-SYNTHETIC): The IHDP dataset has gained popularity in the causal inference literature dealing with heterogeneous treatment effects [1, 18, 28, 43]. The original data includes 25 real covariates and comes from a randomized experiment to evaluate the benefit of IHDP on IQ scores of three-year-old children. A selection bias was introduced by removing some of the treated population, thus resulting in 608 control patients and 139 treated (747 total). The outcomes were simulated using the standard non-linear ‘Response Surface B’ as described in [18].

OPIOID: We sampled a sub-population consisting of healthcare claims for five million patients from the MarketScan Commercial claims database. These claims describe patients’ medical histories, including both inpatient admissions and outpatient services. Diagnoses, procedures, prescriptions and dosages are recorded. We follow the cohort selection procedure outlined in Zhang et al. [50] to filter patients based on several criteria.

For each patient in our final cohort, we create a feature vector that includes basic demographic information such as age, gender and geographic region. We also included a predefined set of procedures along with diagnostic codes which are associated with opioids and/or addiction, based on input from a physician.

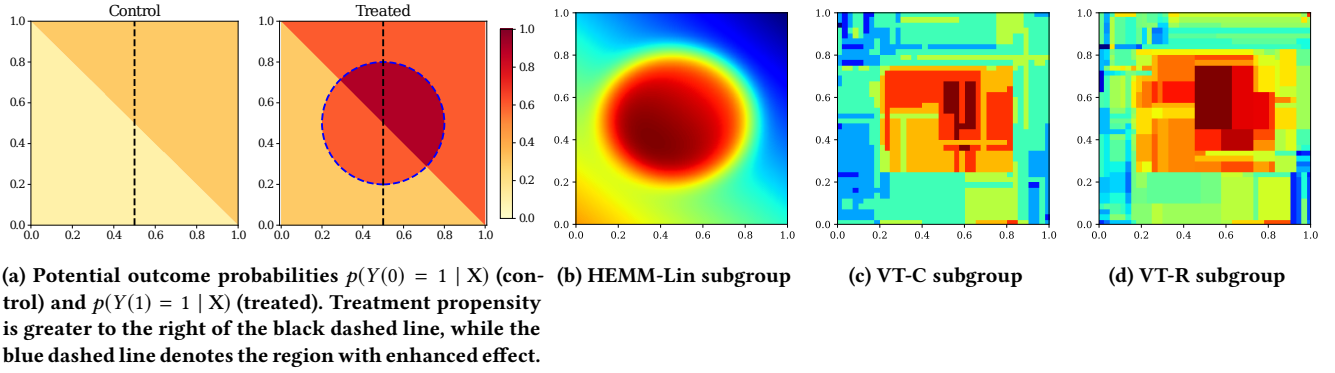


Figure 3: SYNTHETIC dataset and enhanced effect subgroups discovered by HEMM and Virtual Twins (VT).

We label all patients who received addiction diagnoses and patients who continued use of opioids for more than one year after the initial prescription as belonging to the positive (adverse outcome) class. Patients who discontinued opioid use within one year of initial treatment were labeled as negative. We use the terms “addicted” and “not addicted” as shorthand for these outcomes. Patients prescribed natural or semi-synthetic opioids are considered the control group, whereas patients administered synthetic opioids are considered the treated group. Table 1 summarizes the basic statistics of this dataset.

6 EXPERIMENTAL RESULTS

We evaluated the proposed HEMM quantitatively on two tasks, prediction of heterogeneous treatment effects and identification of subgroups with enhanced or reduced effect. These results are discussed in Sections 6.2 and 6.3 respectively in comparison to existing methods, focusing on those that also estimate heterogeneous effects in an interpretable manner. Methods used in the comparison are described in Section 6.1 and parameter selection details are in Appendix D. In Section 6.4, we provide qualitative results for the **OPIOID** dataset on the features the model discovers as characteristics of “at-risk” individuals, i.e. those in enhanced effect subgroups.

6.1 Algorithms in Comparison

We have considered Virtual Twins (VT) [12], QUINT [10], and SIDES [27] among methods that identify subgroups with different treatment effects. We implemented two versions of VT in which the treatment effect is modeled by a decision tree classifier (VT-C) or regressor (VT-R). For VT-C, to better represent the continuous-valued treatment effect (which is a difference in probabilities even if Y is binary), we use a collection of decision tree classifiers obtained by applying different thresholds to the treatment effect.

For QUINT and SIDES, we utilized the standard R implementations and performed extensive hyperparameter tuning. However both QUINT and SIDES failed to recover any subgroups on **SYNTHETIC** and **OPIOID** and we thus did not consider them further. For QUINT, the likely reason is that its assumption of a subgroup with diminished effect is not always met, whereas for SIDES, there may be a numerical issue in how it discretizes continuous covariates.

In terms of methods that only predict heterogeneous effects and do not identify subgroups, we also compare our method with some common approaches in Table 2. Here Linear-1 corresponds to a single ordinary least squares (for continuous outcomes) or logistic regression (for binary outcomes) for both factual and counterfactual outcomes. In the case of Linear-2, we fit two separate linear models to the control and treated populations to better accommodate selection bias and confounding. The other baselines, k -NN, GP, and CFRF are non-parametric versions of this approach where the estimators of the factual and counterfactual outcomes are k -nearest neighbours, Gaussian processes with a linear kernel and Random Forests respectively.

6.2 Heterogeneous Effect Estimation

We first evaluate our performance on estimation of the CATE ($\mathbb{E}[Y(1) - Y(0) | X]$). A popular metric for this evaluation is the *Precision in Estimating Heterogeneous Effects* (PEHE). The PEHE is defined as

$$\text{PEHE} = \frac{1}{n} \sum_{i=1}^n (f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \mathbb{E}[Y(1) - Y(0) | X = \mathbf{x}_i])^2.$$

Here $f_1(\cdot)$ and $f_0(\cdot)$ are the estimated potential outcomes under treatment and control, respectively.

Table 2 compares the performance of HEMM against the methods described in Section 6.1 on both in-sample PEHE (corresponding to a retrospective study) computed on the training data, and out-of-sample PEHE computed on held-out test data. HEMM-MLP and HEMM-Lin refer to the proposed approach with f in (6), (7) as a multilayer perceptron and linear function respectively to model the effect of confounders on the outcome.

HEMM consistently outperforms these standard causal inference baselines. GP and Linear-2 perform close to HEMM on **SYNTHETIC**. We noticed that when a larger sample of data points is available to VT-R, its performance increases dramatically. However, its performance drops in higher-dimensional settings as in the case of **IHDP** and **OPIOID**; this is expected with methods involving non-parametric regression.

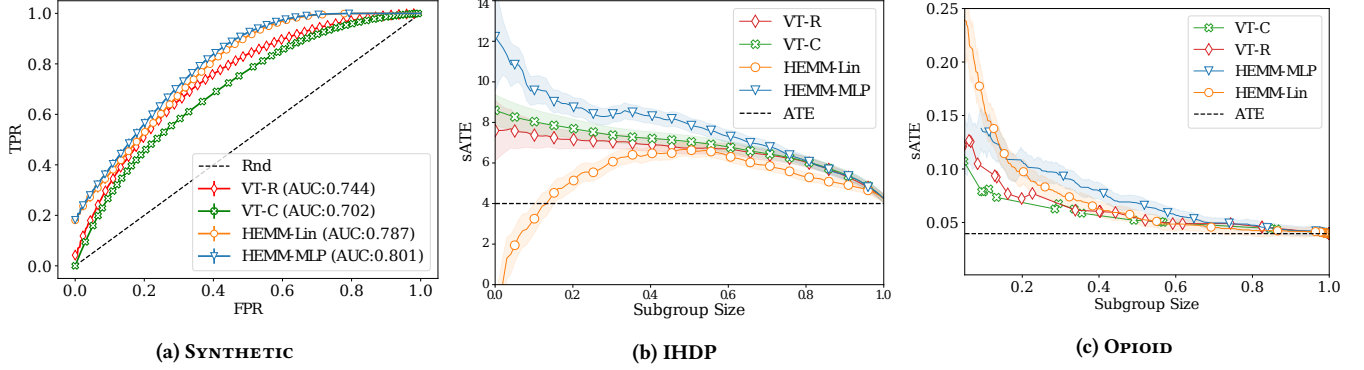


Figure 4: Performance of the proposed HEMM and Virtual Twins on the subgroup discovery task. For SYNTHETIC data, we have access to ground truth labels for the enhanced treatment effect group and hence compare performance using the ROC. For the IHDP and OPIOID datasets, we compare average treatment effect (ATE) estimates within the identified subgroup as a function of subgroup size (as a fraction of the population).

	SYNTHETIC		IHDP	
	In-sample	Out-Sample	In-sample	Out-Sample
HEMM-MLP	0.101 ± 10^{-3}	0.102 ± 10^{-3}	1.6 ± 0.10	1.8 ± 0.10
HEMM-Lin	0.116 ± 10^{-3}	0.116 ± 10^{-3}	2.8 ± 0.32	2.9 ± 0.33
Linear-1	0.278 ± 10^{-3}	0.278 ± 10^{-3}	7.9 ± 0.46	7.9 ± 0.47
Linear-2	0.106 ± 10^{-3}	0.107 ± 10^{-3}	2.3 ± 0.18	2.4 ± 0.21
k-NN	0.210 ± 10^{-3}	0.210 ± 10^{-3}	3.2 ± 0.12	4.2 ± 0.22
GP	0.106 ± 10^{-3}	0.107 ± 10^{-3}	2.1 ± 0.11	2.3 ± 0.14
CFRF	0.146 ± 10^{-3}	0.142 ± 10^{-3}	2.7 ± 0.31	3.3 ± 0.72
VT-R	0.130 ± 10^{-3}	0.130 ± 10^{-3}	2.5 ± 0.26	2.9 ± 0.51

Table 2: $\sqrt{\text{PEHE}}$ values in estimating heterogeneous effects. Error represents 95% confidence interval of multiple Monte Carlo initializations.

6.3 Subgroup Identification

SYNTHETIC: In the case of synthetic data, the subgroup with enhanced treatment effect is known. We first visualize the performance of HEMM and VT in identifying this subgroup. For HEMM-Lin, Figure 3b shows the estimated probability $p(Z | X)$ of belonging to the enhanced effect subgroup evaluated on the test set. The true circular region is recovered well. Figures 3c and 3d plot the VT-C and VT-R predictions of CATE on the test set. For VT-C, the prediction represents an average over the collection of decision tree classifiers, while for VT-R, it is simply the output of the decision tree regressor. While the difficulty in reproducing the circular shape is expected for decision trees, the enhanced effect estimates are also less uniform than in Figure 3b.

We also evaluate subgroup identification more quantitatively by treating it as a problem of classifying whether or not points in the test set belong to the enhanced effect subgroup. ROC curves may then be plotted as in Figure 4a. For HEMM, the ROC is traced by varying the threshold on the probability $p(Z | X)$ of being in the enhanced effect subgroup (shown in Figure 3b). Similarly for VT-C and VT-R, the threshold on the CATE estimates (Figures 3c and 3d) is varied. HEMM has higher ROCs than VT (on this example, in

line with Figures 3b–3d. There is little difference between HEMM-Lin and HEMM-MLP since the dependence on the covariates X in Figure 3a is simple and complex adjustment is not needed.

IHDP and OPIOID: For the other two datasets, we conduct a relative comparison with VT since we lack data on ground truth subgroups. The evaluation involves two steps. First, we assign individuals to an enhanced effect subgroup of varying size. (The same procedure can be used for a diminished effect subgroup but we omit the results due to space.) For HEMM, we choose the subgroup k with the largest main effect γ_k and vary the threshold applied to the corresponding membership probability $p(Z = k | X)$ returned by the model. For VT-C and VT-R, we vary the threshold applied to the CATE estimates, either the composite estimate of the decision tree classifiers or the regressor estimate, the same quantities as for the synthetic data.

In the second step, we build a propensity score model (an estimator of treatment propensity $p(T = 1 | X)$) to estimate the average treatment effect (ATE) conditioned on belonging to the enhanced treatment effect subgroup defined in the first step. For the propensity score model $e(X)$, we fit a random forest, for which parameter tuning is performed on the DEV set. We then use the inverse probability of treatment weighting (IPTW) estimator [19] of the ATE within a subgroup S as follows:

$$\hat{\tau}_S = \frac{1}{|S|} \sum_{i \in S} \left(\frac{y_i t_i}{e(X_i)} - \frac{y_i (1 - t_i)}{1 - e(X_i)} \right). \quad (13)$$

IPTW estimation is used for both HEMM- and VT-defined subgroups to be consistent.

Figures 4b and 4c plot subgroup ATE versus subgroup size (as a fraction of the population) as the threshold for subgroup assignment is varied. When the subgroup is the entire population at size 1.0, all curves meet at the population ATE (dashed line). Since we have selected the enhanced effect subgroup, the curves are then expected to increase as the subgroup is restricted to individuals with larger treatment effects. The fact that this increase is nearly monotonic for HEMM-MLP is evidence for the validity of the discovered subgroup, since the IPTW estimator used here is an independent check on the

Musculoskeletal System	Nervous System
1.0 spinal curve (kyphosis, lordosis, scoliosis)	1.0 extrapyramidal diseases/movmt. disorders
1.0 ankle fracture	1.0 idiopathic peripheral neuropathies
1.0 sprains/strains of hand and wrist	1.0 headaches
Integumentary System	Endocrine System
1.0 cellulitis and abscess of finger and toe	.70 simple and unspecified goiter
1.0 local skin infections	.67 other endocrine disorders
1.0 psoriasis and similar disorders	.65 thyrotoxicosis with or without goiter
Reproductive System	Digestive and Excretory Systems
1.0 female infertility	.71 benign neoplasm of intestinal tract
.82 testicular dysfunction	.69 inguinal hernia
.82 disorders of penis	.58 diverticulitis
Circulatory System	Immune System
1.0 hypertensive heart disease	.56 immunization
.79 other disorders of circulatory system	.54 strep throat and scarlet fever
.70 cardiac dysrhythmias	.52 bacterial infections in other conditions
Nutrition	Visual System
1.0 BMI	.87 keratitis
1.0 b-complex deficiency	.60 other disorders of eye (epi)scleritis
.71 disorder of electrolyte/acid-base balance	.57 visual disturbances
Auditory System	Psychology
.53 vertiginous syndrome/vestibular disorder	1.0 suspected mental health condition
.50 otitis media/eustachian tube disorders	.58 adjustment reaction
.44 disorders of pinna and mastoid process	.55 nondependent abuse of drugs
Digestive System (upper/oral)	Respiratory System
.77 hernia, abdominal cavity w/o obstruction	1.0 other diseases of respiratory tract
.77 dentofacial anomalies of jaw	.74 deviated nasal septum
.76 diseases of oral soft tissues	.69 influenza

Table 3: Top features of the enhanced effect subgroup k discovered by HEMM-MLP on the OPIOID dataset. The numbers are the ratios $\pi_{jk}/\sum_{k'} \pi_{jk'}$, where 1/2 represents no increase in prevalence over the other subgroup ($K = 2$).

treatment effect model (6), (7) used by HEMM. Compared to VT, the subgroups identified by HEMM-MLP have higher ATE. This suggests that for a given subgroup size, HEMM-MLP is better at grouping together individuals with more enhanced effects. HEMM-Lin on the other hand displays contrasting performances. On IHDP in Figure 4b, the estimated ATE actually decreases for subgroup sizes less than 0.5, likely due to the inadequacy of a linear model to adjust for confounding and accurately estimate CATE. In Figure 4c however, the ATE does increase monotonically and faster than for VT.

6.4 Interpretation of the OPIOID Enhanced Effect Subgroup

We now turn our focus to the motivating application of opioids and analyze key characteristics of the enhanced effect subgroup, i.e. those patients at greater risk of adverse outcomes when treated initially with synthetic opioids. To interpret these features, we collaborated with a subject matter expert (SME) with a PhD in cognitive neuroscience and a clinical research emphasis in chronic pain conditions and treatments, including opioids. Table 3 shows the top features of the enhanced effect subgroup as identified by HEMM-MLP. The features are organized by general bodily system and sorted in descending order of prevalence relative to the other subgroup (see table caption); the selection of 3 features in each system was arbitrary and chosen primarily for simplicity and space constraints.

Patients with a history of chronic conditions in general, as well as chronic pain conditions more specifically, are at an increased risk for addiction. Many of the chronic conditions in Table 3, e.g. heart

disease (circulatory system), psoriasis (integumentary system), and BMI/obesity (nutrition) also appear in the CDC opioid prescribing guidelines [9] or have extensive literature linking them to increased risk for long-term pain, either intrinsic to the condition or due to needed medical procedures that are more likely to expose patients to opioids [14]. For example, numerous papers show a link between increased body-mass index (BMI) and increased pain intensity and duration (with anti-correlations between BMI and pain recovery) [33], and obesity has also been associated with higher initial opioid doses [25]. Additionally, the chronic nutritional deficiencies and imbalances shown in Table 3 have been linked to acute but intense muscle spasms as well as peripheral polyneuropathies and paresthesias (see e.g. [31]), pain disorders which also show up as increased risk factors (nervous system).

Regarding chronic pain conditions, patients with a history of abnormal spinal curvatures (which can produce low back pain and neuropathy), idiopathic peripheral neuropathies, and headaches (musculoskeletal and nervous systems) are at increased risk for addiction. These are not surprising as they are notoriously difficult to treat using non-opioid therapies such as non-steroidal anti-inflammatory drugs (NSAIDs), steroids, or common procedures and surgeries (e.g. joint replacement or local injections) [7]. They involve pain that may be severely intense or debilitating, sometimes unpredictable or idiopathic, and often non-specific or diffuse (pain is referred, not well localized, or difficult to describe) and thus require a cocktail of prescription medications or invasive procedures, increasing the likelihood of exposure to opioids [36, 39, 47]. The intensity, duration, and non-specificity of pain may also be a reason why digestive excretory, digestive (upper/oral), and reproductive conditions also show up as moderately strong features in Table 3. These diagnoses may either directly result in acute or chronic non-somatic visceral pain (hernia, diverticulitis) [8] or relate to conditions with chronic visceral pain (e.g. female infertility may be secondary to endometriosis or pelvic inflammatory diseases). Opioids are widely utilized for such visceral pain conditions [13], although often in short duration due to adverse events. Notably, some minor injuries causing acute or procedure-related pain (ankle fractures or hand/wrist sprains – musculoskeletal system) also feature prominently, as do features related to the mouth (dentofacial abnormalities and soft oral tissues – digestive system (upper/oral)). This is also expected given that opioids are most commonly prescribed for post-surgical or intense acute pain [4, 15]; regarding dental procedures specifically, previous research suggests that a substantial proportion of adults are first exposed to opioids through dental procedures [42].

Another expected finding was that individuals with psychological comorbidities (mental health conditions) also have high probability of belonging to the enhanced response group, with individuals participating in psychotherapy having reduced risk of addiction (psychotherapy was among the lowest-scored features and hence not shown in the table). Substantial research has already linked mental illness with opioid misuse [14, 39, 47]. Although adjustment reaction (psychology) appears with a lower score in Table 3, it encompasses reactions to trauma, episodic emotional disorders, and chronic anxiety that have been shown to be comorbid with many of the chronic diagnoses and pain conditions discussed above [14, 39, 47] and have also been linked with increased opioid dosages

[16]. Similarly, nondependent abuse of drugs also has a lower score but it is well known that opioid dependence and addiction are associated with polysubstance use and abuse [38, 44].

A small subset of features with high scores were more challenging to interpret, likely because they pick up on subtle relationships between existing clinical variables or hidden variables. For example, electrolyte imbalances (nutrition) occur often and in many situations, making it difficult to speculate on why they were a top feature for risk propensity. Likewise, skin infections and abscesses (integumentary system) also are common and non-specific. However, it is possible that these features are secondary symptoms of important risk features. For instance, electrolyte imbalances are commonly seen in alcoholism and substance use disorders [34], as are skin infections and abscesses [45].

Based on this initial overview, the SME judged the majority of the identified features to be scientifically meaningful with potential clinical utility for future prescribing guidelines. In summary, acute or chronic conditions that put patients at increased risk for initial exposure to opioids, via acute procedures or comorbid prolonged intense pain, both increased a patient's addiction likelihood. Comorbid mental health disorders, particularly those related to stress or trauma and substance abuse also put individuals at greater risk for future opioid addiction.

7 CONCLUSION

We presented a Heterogeneous Effect Mixture Model (HEMM) for inferring subgroups of individuals that exhibit an enhanced effect caused by treatment. Our work contrasts with existing heterogeneous effect estimation methods as we learn interpretable subgroups using soft assignments while retaining expressiveness in the model. The latter is attributed to the capabilities of neural networks, used here to adjust for confounding. We evaluated the performance of HEMM on a synthetic dataset, the semi-synthetic IHDP dataset, and a large real-world healthcare claims dataset (**OPIOID**).

We additionally conducted qualitative analysis of the results obtained by HEMM on the **OPIOID** dataset. Some of our findings are in accordance with existing CDC opioid prescribing guidelines. However, our interpretations are preliminary and future analyses are needed to better understand these features and their relationships. A longer-term goal is to translate such insights into a policy white paper on data-driven, causally-valid opioid prescribing guidelines.

ACKNOWLEDGMENTS

This work was conducted under the auspices of the IBM Science for Social Good initiative. The authors thank Ching-Hua Chen, Fredrik Johansson, Aleksandra Mojsilović, Peder Olsen, Jinghe Zhang, and colleagues at IBM Watson Health for assistance.

REFERENCES

- [1] Ahmed M Alaa and Mihaela van der Schaar. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*. 3424–3432.
- [2] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 112, 518 (2017), 859–877.
- [3] Gabriel A. Brat, Denis Agniel, Andrew Beam, Brian Yorkgitis, Mark Bicket, Mark Homer, Kathe P. Fox, Daniel B. Knecht, Cheryl N McMahon-Walraven, Nathan Palmer, and Isaac Kohane. 2018. Postsurgical Prescriptions for Opioid Naïve Patients and Association with Overdose and Misuse: Retrospective Cohort Study. *BMJ* 360 (Jan. 2018), j5790.
- [4] Chad M. Brummett, Jennifer F. Waljee, Jenna Goesling, Stephanie Moser, Paul Lin, Michael J. Englesbe, Amy S. B. Bohnert, Sachin Kheterpal, and Brahmajee K. Nallamothu. 2017. New Persistent Opioid Use After Minor and Major Surgical Procedures in US Adults. *JAMA Surg.* 152, 6 (2017), e170504.
- [5] Robert M Califf, Janet Woodcock, and Stephen Ostroff. 2016. A proactive response to prescription opioid abuse. *New England J. Med.* 374, 15 (2016), 1480–1485.
- [6] Zhengping Che, Jennifer St. Sauver, Hongfang Liu, and Yan Liu. 2017. Deep Learning Solutions for Classifying Patients on Opioid Use. In *AMIA Annu. Symp. Proc.* 525–534.
- [7] Leslie J. Crofford. 2013. Use of NSAIDs in Treating Patients with Arthritis. *Arthritis Res. Ther.* 15, Suppl. 3 (July 2013), S2.
- [8] Mellar P. Davis. 2012. Drug Management of Visceral Pain: Concepts from Basic Research. *Pain Res. Treat.* 2012 (2012), 265605.
- [9] Deborah Dowell, Tamara M Haegerich, and Roger Chou. 2016. CDC guideline for prescribing opioids for chronic pain—United States, 2016. *JAMA* 315, 15 (2016), 1624–1645.
- [10] Elise Dusseldorp and Iven Van Mechelen. 2014. Qualitative Interaction Trees: A Tool to Identify Qualitative Treatment–Subgroup Interactions. *Stat. Med.* 33, 2 (Jan. 2014), 219–237.
- [11] Mark J. Edlund, Diane Steffick, Teresa Hudson, Katherine M. Harris, and Mark Sullivan. 2007. Risk Factors for Clinically Recognized Opioid Abuse and Dependence among Veterans using Opioids for Chronic Non-Cancer Pain. *Pain* 129, 3 (June 2007), 355–362.
- [12] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. 2011. Subgroup identification from randomized clinical trial data. *Stat. Med.* 30, 24 (2011), 2867–2880.
- [13] G. F. Gebhart, Xin Su, Shailen Joshi, N. Ozaki, and J. N. Sengupta. 2000. Peripheral Opioid Modulation of Visceral Pain. *Ann. N. Y. Acad. Sci.* 909, 1 (Jan. 2000), 41–50.
- [14] Jason M. Glanz, Komal J. Narwaney, Shane R. Mueller, Edward M. Gardner, Susan L. Calcaterra, Stanley Xu, Kristin Breslin, and Ingrid A. Binswanger. 2018. Prediction Model for Two-Year Risk of Opioid Overdose Among Patients Prescribed Chronic Opioid Therapy. *J. Gen. Intern. Med.* 33, 10 (Oct. 2018), 1646–1653.
- [15] Calista M. Harbaugh, Jay S. Lee, Hsueh Mei Hu, Sean Esteban McCabe, Terri Voepel-Lewis, Michael J. Englesbe, Chad M. Brummett, and Jennifer F. Waljee. 2018. Persistent Opioid Use Among Pediatric Patients After Surgery. *Pediatrics* 144, 1 (Jan. 2018), e20172439.
- [16] G. T. T. Helmerhorst, A.-M. Vranceanu, M. Vrahas, M. Smith, and D. Ring. 2014. Risk Factors for Continued Opioid Use One to Two Months After Surgery for Musculoskeletal Trauma. *J. Bone & Joint Surgery* 96, 6 (March 2014), 495–499.
- [17] Miguel A. Hernán and James M. Robins. 2018. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, USA. Forthcoming.
- [18] Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.
- [19] Guido W Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86, 1 (2004), 4–29.
- [20] Anupam B. Jena, Dana Goldman, Leonard D Schaeffer, Lesley Weaver, and Pinar Karaca-Mandic. 2014. Opioid Prescribing by Multiple Providers in Medicare: Retrospective Observational Study of Insurance Claims. *BMJ* 348 (Feb. 2014), g1393.
- [21] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *Proc. Int. Conf. Mach. Learn.* 3020–3029.
- [22] Nayoung Kim, Jonas L. Matzon, Jack Abboudi, Christopher Jones, William Kirkpatrick, Charles F. Leinberry, Frederic E. Liss, Kevin F. Lutsky, Mark L. Wang, Mitchell Maltenfort, and Asif M. Ilyas. 2016. A Prospective Evaluation of Opioid Utilization After Upper-Extremity Surgical Procedures: Identifying Consumption Patterns and Determining Prescribing Guidelines. *J. Bone Joint Surg.* 98, 20 (Oct. 2016), e89.
- [23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [24] Michael P. Klueh, Hsueh M. Hu, Ryan A. Howard, Joceline V. Vu, Calista M. Harbaugh, Pooja A. Lagisetty, Chad M. Brummett, Michael J. Englesbe, Jennifer F. Waljee, and Jay S. Lee. 2018. Transitions of Care for Postoperative Opioid Prescribing in Previously Opioid-Naïve Patients in the USA: a Retrospective Review. *J. Gen. Intern. Med.* in press (2018).
- [25] A. M. Kobus, D. H. Smith, B. J. Morasco, E. S. Johnson, X. Yang, A. F. Petrik, and R. A. Deyo. 2012. Correlates of higher-dose opioid medication use for low back pain in primary care. *J. Pain* 13, 11 (Nov. 2012), 1131–1138.
- [26] Anders Krogh and John A. Hertz. 1992. A simple weight decay can improve generalization. In *Advances in neural information processing systems*. 950–957.
- [27] Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas. 2011. Subgroup identification based on differential effect search — a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat. Med.* 30, 21 (2011), 2601–2621.
- [28] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Adv. Neur. Inf. Process. Syst.* 6446–6456.

- [29] Martin A Makary, Heidi N Overton, and Peiqi Wang. 2017. Overprescribing is major contributor to opioid crisis. *BMJ* 359 (2017), j4792.
- [30] Benjamin M. Marlin, Mark Schmidt, and Kevin P. Murphy. 2009. Group Sparse Priors for Covariance Estimation. In *Proc. Conf. Uncertainty Artif. Intell.* Montreal, Canada, 383–392.
- [31] Barbara Mostacci, Rocco Liguori, and Arrigo F. Cicero. 2018. Nutraceutical Approach to Peripheral Neuropathies: Evidence from Clinical Trials. *Current Drug Metabolism* 19, 5 (2018), 460–468.
- [32] Daniel B. Neill and William Herlands. 2018. Machine Learning for Drug Overdose Surveillance. *J. Tech. Human Serv.* 36, 1 (Jan. 2018), 8–14.
- [33] Akiko Okifuji and Bradford D. Hare. 2015. The association between chronic pain and obesity. *J. Pain Res.* 8 (2015), 399–408.
- [34] Biff F. Palmer and Deborah J. Clegg. 2017. Electrolyte disturbances in patients with chronic alcohol-use disorder. *New England J. Med.* 377, 14 (2017), 1368–1377.
- [35] Stephen T. Parente, Susan S. Kim, Michael D. Finch, Lisa A. Schloff, Thomas S. Rector, Raafat Seinfeldin, and J. David Haddox. 2004. Identifying Controlled Substance Patterns of Utilization Requiring Evaluation Using Administrative Claims Data. *Am. J. Managed Care* 10, 11 (Nov. 2004), 783–790.
- [36] Pravinkumar R. Patil, Jonathan Wolfe, Qayyim Said, Jeremy Thomas, and Bradley C. Martin. 2015. Opioid Use in the Management of Diabetic Peripheral Neuropathy (DPN) in a Large Commercially Insured Population. *Clin. J. Pain* 31, 5 (May 2015), 414–424.
- [37] Judea Pearl. 2009. *Causality*. Cambridge university press.
- [38] Joseph V. Pergolizzi, Christopher Gharibo, Steven Passik, Sumedha Labhsetwar, Robert Taylor, Jason S. Pergolizzi, and Gerhard Müller-Schwefe. 2012. Dynamic risk factors in the misuse of opioid analgesics. *J. Psychosomatic Res.* 72, 6 (2012), 443–451.
- [39] A. Rosenblum, L. A. Marsch, H. Joseph, and R. K. Portenoy. 2008. Opioids and the treatment of chronic pain: controversies, current status, and future directions. *Exp. Clin. Psychopharmacol.* 16, 5 (Oct. 2008), 405–416.
- [40] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Edu. Psych.* 66, 5 (1974), 688–701.
- [41] Donald B. Rubin. 1986. Which Ifs Have Causal Answers. *J. Am. Stat. Assoc.* 81, 396 (1986), 961–962.
- [42] Alan R. Schroeder, Melody Dehghan, Thomas B. Newman, Jason P. Bentley, and K. T. Park. in press. Association of Opioid Prescriptions From Dental Clinicians for US Adolescents and Young Adults With Subsequent Opioid Use and Abuse. *JAMA Intern Med.* (in press).
- [43] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proc. Int. Conf. Mach. Learn.* Sydney, Australia, 3076–3085.
- [44] M. Soyka. 2015. Alcohol Use Disorders in Opioid Maintenance Therapy: Prevalence, Clinical Correlates and Treatment. *Eur. Addict. Res.* 21, 2 (Jan. 2015), 78–87.
- [45] Sandra A Springer, P Todd Korthuis, and Carlos Del Rio. 2018. Integrating treatment at the intersection of opioid use disorder and infectious disease epidemics in medical settings. *Ann Intern Med* 169, 5 (2018), 335–336.
- [46] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10, Feb (2009), 141–158.
- [47] Nora D. Volkow and A. Thomas McLellan. 2016. Opioid Abuse in Chronic Pain — Misconceptions and Mitigation Strategies. *N. Engl. J. Med.* 374, 13 (2016), 1253–1263.
- [48] Tong Wang and Cynthia Rudin. 2017. Causal Rule Sets for Identifying Subgroups with Enhanced Treatment Effect. arXiv:1710.05426.
- [49] Ming Yuan and Yi Lin. 2006. Model Selection and Estimation in Regression with Grouped Variables. *J. R. Stat. Soc. B* 68, 1 (2006), 49–67.
- [50] Jinghe Zhang, Vijay S. Iyengar, Dennis Wei, Bhanukiran Vinzamuri, Hamsa Bastani, Alexander R. Macalalad, Anne E. Fischer, Gigi Yuen-Reed, Aleksandra Mojsilović, and Kush R. Varshney. 2017. Exploring the Causal Relationships between Initial Opioid Prescriptions and Outcomes. In *AMIA Workshop Data Min. Med. Informat.* Washington, DC, USA.

APPENDIX

A IDENTIFIABILITY

THEOREM 1 (IDENTIFIABILITY). *Under the Directed Acyclic Graph in Figure. 2,*

$$p(Y|\mathbf{do}(T = t), X) = \int_Z p(Y|X, Z, T = t)p(Z|X)$$

Proof.

$$p(Y|\mathbf{do}(T = t), X) = \int_Z p(Y|\mathbf{do}(T = t), Z, X)p(Z|\mathbf{do}(T = t), X) \\ \text{(conditioning on and marginalizing out } Z)$$

Now, $p(Y|\mathbf{do}(T = t), Z, X) = p(Y|T = t, Z, X)$

and, $p(Z|\mathbf{do}(T = t), Z) = p(Z|T = t, X)$

(From [37]’s Backdoor Adjustment Formula)

$$p(Y|\mathbf{do}(T = t), X) = \int_Z p(Y|\mathbf{do}(T = t), Z, X)p(Z|T = t, X)$$

(But, under the DAG assumptions, $Z \perp\!\!\!\perp T|X$)

$$\text{Thus, } p(Y|\mathbf{do}(T = t), X) = \int_Z p(Y|X, Z, T = t)p(Z|X) \quad \blacksquare$$

B PARAMETER INFERENCE WITH EM

In this section we provide an alternate approach to perform parameter inference using Expectation Maximization and compare it to the proposed ELBO optimization.

B.1 Inference

The complete-data log-likelihood used in EM is given by

$$\mathcal{L}_c(\Theta, \mathcal{D}) = \sum_{i=1}^N \sum_{k=1}^K \mathbf{1}\{z_i = k\} \ln(P_k^m(\mathbf{x}_i)P_k^t(y_i)), \quad (14)$$

here, $P_k^m(\mathbf{x}_i) = p(z_i = k|\mathbf{x}_i)$, $P_k^t(y_i) = p(y_i|\mathbf{x}_i, t_i, z_i = k)$ and $\mathbf{1}$ is the indicator function.

E-Step

As is standard in EM, let us define $Q(\Theta, \Theta^l)$ as the expected value of the complete-data log-likelihood (14) with respect to the conditional distribution of the latent variable given the current parameters Θ^l :

$$Q(\Theta, \Theta^l) = \mathbb{E} \left[\mathcal{L}_c(\Theta, \mathcal{D}) \mid \{y_i, \mathbf{x}_i, t_i\}_{i=1}^N, \Theta^l \right].$$

Since the only quantity in (14) that depends explicitly on z_i is the indicator $\mathbf{1}\{z_i = k\}$, we can compute Q by replacing these indicators with the posterior probability of $z_i = k$:

$$\begin{aligned} \mathbb{E}[\mathbf{1}\{z_i = k\}] &\equiv h_i^{(k)} = p(z_i = k|y_i, \mathbf{x}_i, t_i; \Theta^l) \\ &= \frac{p(y_i|z_i = k, \mathbf{x}_i, t_i; \Theta^l)p(z_i = k|\mathbf{x}_i, \Theta^l)}{p(y_i|\mathbf{x}_i, t_i, \Theta^l)} \\ &= \frac{p(y_i|z_i = k, \mathbf{x}_i, t_i; \Theta^l) p(\mathbf{x}_i|z_i = k; \Theta^l)p(z_i = k; \Theta^l)}{p(y_i|\mathbf{x}_i, t_i, \Theta^l) p(\mathbf{x}_i; \Theta^l)}. \end{aligned} \quad (15)$$

The terms in the numerator can be evaluated using (1)–(3), (6) from the model. The terms in the denominator are normalization constants that ensure the probabilities sum to one.

M-Step

We use a gradient ascent method in the M-step to maximize Q with respect to the parameters of the model:

$$\Theta^{l+1} = \arg \max_{\Theta} \left(\sum_{i=1}^N \sum_{k=1}^K h_i^{(k)} \ln[P_k^m(\mathbf{x}_i)P_k^t(y_i)] - \lambda \Omega(\pi) \right). \quad (16)$$

The posterior probabilities $h_i^{(k)}$ are fixed from the E-step. Using Bayes’ rule as in (15), $P_k^m(\mathbf{x}_i) = p(z_i = k|\mathbf{x}_i)$ can be expressed in terms of model parameters μ_k, Σ_k, π_k defined by (1)–(3). Similarly, $P_k^t(y_i)$ depends on parameters w and γ_k according to the outcome

model (6). The use of gradient ascent allows for any differentiable nonlinear function $f(\cdot)$ in (6). Lastly, the log-prior term $-\lambda\Omega(\boldsymbol{\pi})$ favors sparsity in $\boldsymbol{\pi}$ according to either (4) or (5), where λ is a parameter controlling the strength of the prior.

Instead of computing Q over the entire dataset, we sample a mini-batch from the dataset and perform the E-step and M-step over just the mini-batch in each iteration. We observe that this mini-batch procedure is faster than regular EM over the entire dataset.

B.2 Comparison to ELBO Optimization

In order to compare the performance of EM vis-à-vis the variational inference-motivated ELBO optimization, we compare the train and test negative log-likelihood for both approaches on 100 realizations of the **IHDP** dataset, with the number of latent components, $K = 3$ and stochastic gradient descent learning rate of 1×10^{-3} . We then average over the resulting 100 curves. Figure 5 presents the results:

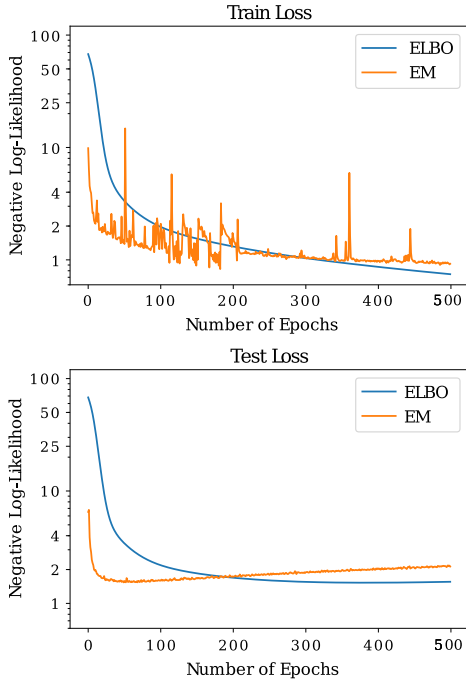


Figure 5: The negative log-likelihood (NLL) versus the number of optimization epochs for EM and ELBO. Notice how the Test NLL continues to decrease for ELBO vs. EM, suggesting the ELBO approach is less sensitive to overfitting.

it is clear from the figure that the ELBO approach has less tendency to overfit and results in an overall better fit compared to the EM approach. This motivates our choice to directly optimize the ELBO.

C PARAMETER INITIALIZATION

Gradient based optimization strategies can be subject to local minima and hence their performance is dependent on parameter initialization. To initialize the model with ‘good’ values ensuring better convergence, we set the mean for each component, $\boldsymbol{\mu}_k$ and $\boldsymbol{\pi}_k$, equal to the sample mean of the entire data, i.e. $\boldsymbol{\mu}_k^0 = \frac{1}{N} \sum_i \mathbf{x}_{\text{cont}, i}$,

$\boldsymbol{\pi}_k^0 = \frac{1}{N} \sum_i \mathbf{x}_{\text{disc}, i}$, and the covariance of every component to $\Sigma_k^0 = \text{diag}(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ is a vector consisting of the sample variances of the continuous covariates \mathbf{X}_{cont} . We pre-train the parameters \mathbf{w}_t in the outcome model (6) using standard cross-entropy loss without the subgroup and treatment assignment term $\gamma_k t$. Finally, we initialize the treatment coefficients, γ_k , randomly with positive values for all k .

D MODEL FITTING

Our implementation of HEMM has two free parameters, the number of groups K and the strength of the sparsity prior, λ . For the **OPIOID** dataset we divide the dataset into 3 parts with 70% as TRAIN for model training, 10% as DEV for parameter tuning, and 20% as TEST for evaluation. The partition is done so that the joint distribution of outcome and treatment is approximately the same in the 3 sets: $p_{\text{TRAIN}}(Y, T) \approx p_{\text{TEST}}(Y, T) \approx p_{\text{DEV}}(Y, T)$.

For **IHDP** we use the standard 80/20 TRAIN/TEST split as is popular in literature. We perform a grid search over $K \in \{2, 3, 4\}$ and $\lambda \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$. For each (K, λ) pair, we perform 5 runs with randomly initialized values of the treatment coefficients γ_k . All other parameters are initialized as described in Section C. For

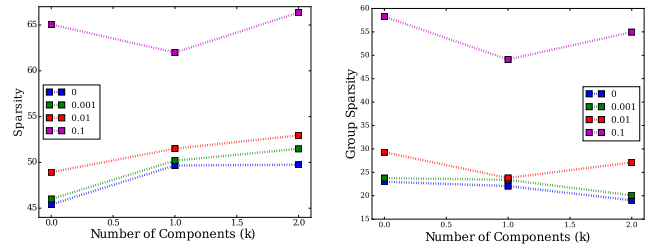


Figure 6: Effect of the prior strength λ for Laplace (ℓ_1 , left) and group $\ell_{1,2}$ (right) priors and different values of K .

Adam, we use a step size of 10^{-4} and mini-batch sizes of 10, 20 & 1000 for **SYNTHETIC**, **IHDP**, and **OPIOID** respectively, and stop parameter update if the ELBO on the DEV is lower at the end of an epoch. We also search over the space of models where the outcome and counterfactual have the same or different parameterisation based on treatment assignment. From all the (K, λ) pairs and random initializations above, we select the model that has the best performance on the DEV set in predicting the outcome y_i , in terms of the Area Under the Receiver Operating Characteristic (AU-ROC).

For the **SYNTHETIC** dataset, we simply set $K = 2$ and $\lambda = 0$. In this case there is no need for a DEV set and the data is split 50/50 between TRAIN and TEST. Figure 6 shows the percentage of parameters π_{jk} equal to zero (labeled “Sparsity” and “Group Sparsity” in the plots) across all groups k for different values of (K, λ) and averaged over random initializations. It is clear that larger values of the prior strength parameter λ result in sparser solutions.