Interpretable (not just posthoc-explainable) heterogeneous survivors bias-corrected treatment effects for assignment of postdischarge interventions to prevent readmissions

Hongjing Xia HONGJING@MEDERRATA.COM
Sound Prediction and *Mederrata Research*

Joshua C. Chang JOSH@MEDERRATA.COM
Mederrata Research and *Sound Prediction*
and *National Institutes of Health, Clinical Center*

Sarah Nowak SARAH.NOWAK@MED.UVM.EDU
University of Vermont

Sonya Mahajan SONYA@MEDERRATA.COM
Rohit Mahajan RO@MEDERRATA.COM
Ted L. Chang TED@MEDERRATA.COM
Sound Prediction and *Mederrata Research*

Carson C. Chow CARSON.CHOW@NIH.GOV
National Institutes of Health, NIDDK and *Mederrata Research*

Abstract

We used survival analysis to quantify the impact of postdischarge evaluation and management (E/M) services in preventing hospital readmission or death. Our approach avoids a common pitfall when applying machine learning to this problem: inflated treatment effect estimates due to survivors bias – where the magnitude of inflation may be conditional on heterogeneous confounders in the population. This bias arises simply because in order to receive an intervention after discharge, a person must not have been readmitted in the intervening period. After deriving an expression for the phantom effect due to survivors bias, we controlled for this and other biases within an inherently interpretable model that quilts together linear functions using Bayesian multilevel modeling. We identified case management services as being the most impactful for reducing readmissions overall.

1. Introduction

For Medicare beneficiaries, almost 20% of hospital discharges are followed by a readmission within 30 days (Jencks et al., 2009). The Centers for Medicare and Medicaid Services (CMS), through the Hospital Readmissions Reduction Program (HRRP), penalizes hospitals for readmission rates above the national average in certain conditions (Khera and Krumholz, 2018; McIlvennan et al., 2015) by reducing their reimbursement rates. States such as Maryland, through its Readmission Reduction Incentive Program (RRIP), also have designed combinations of penalties and incentives around the reduction of hospital readmissions.

Payers and providers alike have mutual interest in seeking low-cost interventions for preempting preventable readmissions. These interventions include discharge planning services such as transfers to less-intensive healthcare institutions, as well as postdischarge outpatient interventions. This paper uses Medicare claims data to quantify the efficacy of postdischarge interventions.

1.1. Evaluation management interventions

Medical claims are a rich longitudinal data source for assessing the efficacy of interventions in individuals on a population-wide scale. Claims consist of billing records that are specific to a patient and provider, recording services and patient-specific health details. Each claim consists of a set of medical billing codes of varying dialects. In the United States, procedures are usually recorded using Current Procedural Terminology (CPT)/Healthcare Common Procedure Coding System (HCPCS) Codes. A subset of these codes known as Evaluation/Management (E/M) encompass services that we wish to study.

E/M codes are divided into subcategories depending on type of service (see Supplemental Methods for HCPCS code ranges). In our dataset there exists greater than one in a thousand incidence of the following postdischarge services: office or inpatient visit, hospital observation, hospital inpatient services, consultation services, nursing facility serves, domiciliary services, home health, prolonged health services, and case management. We restrict our analyses to these broad service categories. Our objective is to estimate the efficacy of these services across heterogeneous patient cohorts present in the data.

The rote usage of common machine learning methods for studying this problem is problematic. Ascertaining the effect of these interventions from observational medical claims is a causal inference problem. To derive valid treatment effects requires adjusting for confounders, an issue common to these problems. *A priori*, one cannot expect that all inpatient episodes have equal probability of being followed by interventions. Hence, one must control for the treatment assignment mechanism.

Additionally, the fundamental problem in analyzing the effect of interventions on readmission is in how the incidence of the outcome itself (readmission or death) censors the interventions. The observation of an intervention implies that the adverse outcome had not occurred before the intervention – hence, the incidence of an intervention becomes a strong predictor for an adverse outcome not occurring. Failure to control for this issue leads to bias in the estimation of effects.

Example 1 (Survival bias due to intervention censorship)
Suppose one is ascertaining the impact of an intervention on the incidence of either readmission or death within 30 days. Suppose also that this intervention has no real effect, and that this intervention is performed exactly one week after discharge.

On a population level, one may naively estimate the effect of the intervention by cross-tabulating the incidence of the intervention against the incidence of readmission or death within 30 days. In a large sample of Medicare discharges (Fig. 1), it is seen that the rate of readmission at 30 days is approximately 17%. However, the rate of readmission within seven days is approximately 7%. By definition, none of the individuals readmitted before day seven would have received the intervention, as depicted in Fig. 2. Conditional on not being readmitted before day seven, the mean readmission rate is hence $\frac{0.17-0.07}{1-0.07} \approx 11\%$. *Hence,*

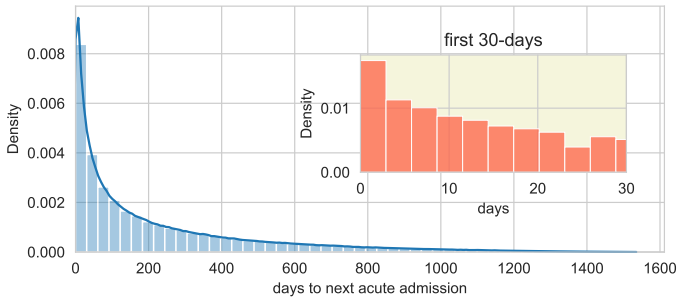

Figure 1: **Empirical density for days post discharge before acute readmission** for a sample of medicare recipients between 2009 and 2012. (inset: the first 30 days). Approximately 17% of discharges are followed by acute readmission within 30 days, and an additional 3% by death.

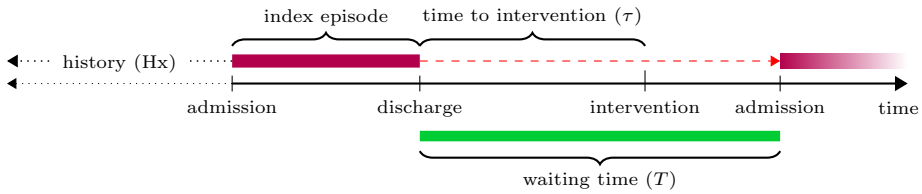

Figure 2: **Waiting time before readmission** and time to intervention for a successfully applied intervention. Only interventions occurring at time $\tau < T$ can be observed. We use Hx to refer to the medical history of the beneficiary prior to admission for the index episode.

this naive cross-tabulation analysis leads to the incorrect conclusion that the ineffective intervention decreases readmission probability from 17% *to* 11%*.*

Classification-based machine learning approaches, which include the intervention as an indicator variable, are susceptible to bias due to this selection issue. Correcting this bias is not as simple as instituting a uniform correction on the effect – the true effect of the intervention can vary and the administration time of the intervention can also vary. Instead, one must explicitly control for this bias.

1.2. Interpretability

The goal of interpretable modeling is to produce predictions that an end-user can understand (Rudin, 2019, 2014; Sudjianto and Zhang, 2021). Medical decision making is high-states and high-risk; knowing how a model makes a decision is crucial to trusting that decision. Although the pairing of black boxes with a particular class of tools known as posthoc explainable AI (xAI) is popular, these tools are no substitute for intrinsic model interpretability. Researchers have consistently shown (Laugel et al., 2019; Kumar et al.,

2020; Slack et al., 2020; Alvarez-Melis and Jaakkola, 2018; Zhou et al., 2022) that these methods are unreliable, often disagreeing with each other (Krishna et al., 2022) and failing to reproduce model ground truth when it is known (Chang et al., 2022).

Instead of relying on explainable-AI, we developed an intrinsically interpretable model that performs as well as black boxes in terms of accuracy. Additionally, we structured our model so that it is mechanistically meaningful – where individual components of the model are themselves interpretable and relatable to real-world observables.

1.3. Generalizable Insights about Machine Learning in the Context of Healthcare

In healthcare, interventions are often subject to a waiting time. Analyzing the impact of these interventions requires controlling for not just heterogeneous rates of treatment assignment, but also the possibility that a target outcome occurs before the intervention may be applied. Failure to do so leads to biased estimates of treatment effects. We introduce methodology for controlling for this bias and producing interpretable heterogeneous local average treatment effects. We applied this methodology to the problem of postdischarge care for preventing all-cause hospital readmissions in Medicare patients and identified cohorts of inpatient episodes that would best benefit from each of several interventions.

2. Related Work

2.1. Readmission modeling

Readmission models in the literature are mostly based on either electronic health records (Golmaei and Luo, 2021; Assaf and Jayousi, 2020; Huang et al., 2020; Liu et al., 2019) or medical claims. Huang et al. (2021) recently provided a survey of readmission modeling efforts, comparing approaches and self-reported performance metrics. In their survey no modeling methodology yielded consistently more-accurate models than others, though some researchers report improvements when using XGBoost or neural networks over over interpretable methods such as logistic regression (Jamei et al., 2017; Liu et al., 2020; Futoma et al., 2015), though not consistently (Allam et al., 2019; Shameer et al., 2016; Min et al., 2019; Larsson et al., 2021; Chang et al., 2022). Reflecting the focus of the HRRP, the literature has focused on 30-day readmissions, though the scope and definition of readmission varies – complicating direct comparisons between studies. Models based on medical claims data typically achieved area under the receiver operator characteristic (AUROC) of approximately 0.7 for predicting their particular 30-day readmission label.

Differences between datasets and their corresponding patient populations also complicate direct comparisons. Our present study is the most-related to two studies using the same dataset. Lahlou et al. (2021) developed an attention-based neural network, reporting an AUROC value of 0.81. However, their outcome label did not distinguish between transfers, planned admissions, and acute admissions in their outcome label so they solve a different problem that is of less practical utility. Our present work shares more similarities with MacKay et al. (2021) , who developed XGBoost models for predicting a set of adverse events, reporting an AUROC of 0.73 for all-cause readmission classification.

2.2. Evaluating postdischarge interventions

In the literature, there exists evidence that postdischarge interventions matter. Broadly, Griffith et al. (2022) found associations between the local availability of postdischarge care options and readmission rates, but their population-level data could not identify specific causal mechanisms that might underlie these relationships. DeLia et al. (2014) investigated the likelihood that a patient would have a follow-up visit after discharge, before being readmitted, and found racial disparities.

Several studies have focused on quantifying the effect of specific interventions – often failing to control for survivors bias. Bricard and Or (2018) studied the impact of follow-up visits for heart failure patients in France, using an instrumental variables approach to control for selection biases in the receipt of postdischarge care – yet did not correct for survivors bias in their analysis. Anderson et al. (2022) studied readmission risk using a Cox-proportional hazards survival model, with the intent of quantifying the effect of followup visits within the first seven days after discharge. However, they also did not control for the survivors bias that is the focus of this manuscript. Vernon et al. (2019) compared readmission rates for UK NHS patients for whom attempted contact was made, offering at-home visit. Interpreting the intervention as the specific choice to attempt contact, made at discharge, their analysis does not suffer from the survivors bias. Harrison et al. (2011) processed commercial medical claims, tabulating readmission rates, comparing those who were contacted via phone call within two weeks of discharge versus all others – this analysis was also affected by survivors bias.

3. Methods

In Example 1, we illustrated the survivors bias that plagues much of the research on postdischarge interventions. The apparent observed treatment effect in these cases is coupled with the statistics of when the intervention was administered in the data, and also the statistics of the waiting time to readmission or death. These couplings induce a phantom effect against which prior studies that did not control for survivors bias should be interpreted. We first derive this effect.

Theorem 1 (Phantom effect due to survivors bias) *Suppose that an intervention administered at time $\tau \sim h(\tau)$ has no effect, then the apparent cumulative readmission probability by time c for cases where interventions have been observed is*

$$\Pr(T \leq c | T \geq \tau) = \int_0^c \left(S_\infty(\tau) - S_\infty(c)\right) h(\tau) \mathrm{d}\tau, \tag{1}$$

where $S_\tau(t) = \int_t^\infty f_\tau(t)\mathrm{d}t$ is the survival function corresponding to the wait time distribution probability density function f_τ, until readmission conditional on receiving an intervention at time τ (f_∞ is the pre-treatment wait time distribution probability density function).

See proof in Appendix A.

Theorem 1 succinctly explains why approaches to ascertaining the effect of such interventions based on two-group comparisons of the readmission probability fail. In order to estimate a valid effect, one must decouple the statistics of the intervention wait from the

statistics of the readmission wait, by taking the time-dependence of interventions into account. To this end, we derive a general expression for the observed data likelihood function that correctly controls for intervention timing.

Proposition 2 (Multiple interventions) *Suppose that* $0 < \tau_1 \leq \tau_2 \leq \ldots \leq \tau_N$ *are fixed times for which* N *interventions are scheduled. Denote* f_∞ *the pre-treatment wait time probability density function, and* $g_{\tau_1,\tau_2,\ldots\tau_K,\infty,\ldots}$ *the probability density function for the remaining waiting time after the* K*–th intervention (occurring at time* τ_K*), if the* $K+1$*-st intervention is never applied. Then*

$$T|\tau_1,\tau_2,\ldots,\tau_N \sim f_{\tau_1,\tau_2,\ldots,\tau_N}(T) \tag{2}$$

where

$$f_{\tau_1,\tau_2,\ldots,\tau_N}(T) = \begin{cases} f_\infty(T) & T < \tau_1 \\ g_{\tau_1,\ldots,\tau_n,\infty,\ldots}(T-\tau_n) & \\ \quad \times \left(1 - \int_0^{\tau_1} f_\infty(T)\mathrm{d}T - \sum_{k=1}^{n} \int_{\tau_k}^{\tau_{k+1}} g_{\tau_1,\ldots,\tau_k,\infty}(T-\tau_k)\mathrm{d}T\right) & T \in (\tau_n, \tau_{n+1}]. \end{cases} \tag{3}$$

See proof in Appendix A.

The estimation of a model as defined as in Eq. 10 is a survival estimation problem.

3.1. Survival

Survival analysis is a collection of statistical methods concerned with characterizing the properties of the wait time to an event. It is particularly suited to problems where some observations are censored, for instance by the closure of a finite observation period. In the presence of right-censoring, one may specify a log-likelihood taking the form

$$\log \pi(\mathbf{T}|\mathbf{\Theta}) = \sum_{n=1}^{N} \left[1_{n,obs} \log \pi_n(T_n|\mathbf{\Theta}) + (1 - 1_{n,obs}) \log \int_{T_n}^{\infty} \pi_n(T|\mathbf{\Theta})\mathrm{d}T\right], \tag{4}$$

where $\mathbf{T} = \{T_n\}_{n=1}^N$ constitutes a set of N independent observations of the wait time of N index inpatient episodes, $\mathbf{\Theta}$ refers to the collection of model parameters that determine the statistics of $\mathbf{T}$, $1_{n,obs}$ is an indicator for whether the event was observed at time T_n (as opposed to censored), and π_n is the predictive density implied by the model for observation n.

As shown in Eq. 3, specification of each π_n requires the specification of a corresponding pre-treatment wait-time density f_∞ as well as the post-treatment wait-time densities $g_{\tau_1,\tau_2,\ldots}$. To do so, it is convenient to model the process from the perspective of hazard functions $\lambda(t) : \mathbb{R}^+ \to \mathbb{R}^+$, which define the time-dependent instantaneous rate of event occurrence. Given a hazard function, one can define a corresponding probability density function

$$f(t) = \lambda(t)e^{-\Lambda(t)}, \qquad \text{where} \qquad \Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u \tag{5}$$

is known as the cumulative hazards function. In Eq. 5, it is evident that the relationship between hazard functions and wait time probability density functions is a bijection. Using hazard functions that change upon intervention, we can capture the behavior of Eq. 3 without explicitly writing out each post-treatment waiting time density. A natural class of models well-suited for this exercise are piece-wise exponential survival regression models (PEM) (Kitchin et al., 1983; Malla and Mukerjee, 2010; Friedman, 1982; Huang et al., 1998).

PEMs are defined through specification of their piece-wise constant hazard functions. PEMs are highly expressive in their capture of time-dependence of survival functions, being able to approximate non-parametric models such as Kaplan-Meier curves (Pepe and Fleming, 1991; Kim and Proschan, 1991; Heuser et al., 2018) using fewer degrees of freedom. In this manuscript, we use the guidance of Chang et al. (2022) in setting breakpoints between time intervals at 1 week, 4 weeks, and 9 weeks after discharge. For each index inpatient episode n, we model the corresponding pre-treatment baseline wait time distribution by specifying the log-hazard within each time interval i,

$$\log \lambda^0_{ni} = \alpha_{ni} + \sum_j \beta_{nj} x_{nj}. \tag{6}$$

In our model, the total log hazard for episode n is a combination of the baseline term, the effect of interventions, and a statistical adjustment to control for bias in treatment assignment (Rubin, 2006; Rhodes, 2010; Bafumi and Gelman, 2007; Feller et al., 2016; Raudenbush et al., 2012). To control for this bias, we jointly model the outcome of postdischarge service assignment, using Poisson regression on pre-treatment variables. We then use this secondary outcome as a predictor in the wait time estimate. Our overall generative model follows

$$\begin{aligned}
I_{nk}|\mu_{nk} &\sim \text{Poisson}(\mu_{nk}) \\
\log \mu_{nk}(t) &= \nu_{nk} + \sum_j \xi_{nkj} x_{nj} \\
\log \lambda_n(t) &= \underbrace{\log \lambda^0_{ni}}_{\text{baseline}} + \underbrace{\sum_{k=1}^{N_n} 1_{t>\tau_{n,k}} \gamma_{n,\tau_k}}_{\text{treatment}} + \underbrace{\sum_{k=1}^{K} \eta_{nk}\mu_{nk}}_{\text{selection adjustment}} \\
T_n|\lambda_n(t) &\sim \text{PEM}(\lambda_n(t))
\end{aligned} \tag{7}$$

where γ_{n,τ_k} is the effect of the intervention administered at τ_k specific to the characteristics of episode n, μ_{nk} is the rate of postdischarge intervention k, predicting I_{nk}, the time-normalized number of interventions for episode n of type k, η_{nk} is an adjustment term to control for bias in the assignment of interventions, and ν_{nk}, ξ_{nk} are parameters of the intervention prediction regression problem.

Note that all parameters in Eq. 7 have an explicit n dependence. Like in hierarchical or mixed effects models, we allow all slopes and intercepts to vary across the dataset. Specifically, we allow these parameters to vary in a piecewise fashion across regions in the dataset, formulating the problem as a hierarchical variant of variable coefficient regression modeling (Hastie and Tibshirani, 1993; Fan and Zhang, 2008; Li et al., 2021).

3.2. Hierarchical modeling for varying coefficients

In this study we adapt the methodology of Chang et al. (2022), utilizing a combination unsupervised methods (Chang et al., 2021) for overlaying a multi-way contingency table over the data and an additive parameter decomposition to vary model parameters between the resulting dataset regions. Effectively, this methodology quilts together inter-related linear generalized linear regression models into a large nonlinear model.

As in that study, we group episodes based on the beneficiary's recent (past year on a quarterly basis) medical utilization history at the time of index admission. We used their pre-trained five-dimensional quarterly utilization embedding (Supplemental Fig. 7), binning each dimension into high and low utilization groups (based on a median cutoff) creating a set of $2^5 = 32$ groupings. Additionally, we included interactions between the history groups with other discrete attributes such as the major diagnostic category (MDC), complication or comorbidity (CC) or a major complication or comorbidity (MCC), whether the length of stay is zero days, whether the primary diagnosis for the admission is acute, and discharge status, to create a high dimensional discrete lattice where the cells define coarse interaction cohorts in the data.

Our objective was to produce a model for each distinct lattice coordinate. Fitting such a model disjointly, by dividing the data, invites overfitting. To combat this issue, Chang et al. (2022) introduced a statistical representation for parameters that takes advantage of shrinkage and partial pooling for inherent regularization. Given a multidimensional lattice, they assign for each parameter a value within the lattice by decomposing the value into the form

$$\theta^{(\boldsymbol{\kappa})} = \overbrace{\theta^{(*,*,\ldots,*)}}^{\text{zero order}} + \overbrace{\theta^{(\kappa_1,*,\ldots,*)} + \theta^{(*,\kappa_2,*,\ldots,*)} + \ldots}^{\text{first order}} + \overbrace{\theta^{(\kappa_1,\kappa_2,\ldots,*)} + \theta^{(\kappa_1,*,\kappa_3,*,\ldots,*)} + \ldots}^{\text{second order}} + \text{H.O.T.}, \tag{8}$$

where $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_D)$ is a D dimensional multi-index. Each term in the decomposition corresponds to a given slice of a multi-way contingency table. We used Gaussian priors for each parameter component, where the variance was weighted proportional to the corresponding count in the multi-way contingency table divided by the total dataset size. This weighting encourages shrinkage of higher-order terms, inducing partial pooling that regularizes the overall solution. For the regression coefficients, we additionally utilized the regularized horseshoe prior for local-global sparse shrinkage (Ghosh and Doshi-Velez, 2017; Bhadra et al., 2019; Polson and Scott, 2011; van Erp et al., 2019). Please see the Supplemental Methods for more details on the model specification.

4. Cohort

For this study we used the CMS Limited Dataset (CMS-LDS) for the years 2008–2013, which was provided as part of the inaugural CMS AI Health Innovations Challenge. This dataset consists of a national 5% beneficiary sample of Medicare fee for service Part A (institutional) and B (outpatient/provider) claims. The 2008 claims had only quarter date specificity so we used them solely to fill out the medical history for 2009 inpatient stays, after assuming that each 2008 claim fell in the middle of its given quarter. We trained the

readmission models on 2009 – 2011 index admissions, and evaluated the models on 2012 index admissions.

Medical claims, generated for billing purposes, require reorganization in order to identify hospital stays. We performed this reorganization by grouping claims based on date, provider, and beneficiary overlap, deriving inpatient episodes of care. After grouping, we filtered, retaining only episodes where the beneficiary had a continuous prior year of Part A/B enrollment. We also excluded episodes from consideration as index episodes if they did not correspond to discharges to less-intensive care (excluding discharge due to death and transfer between facilities of the same acuity). Additionally, we used the official CMS methodology for determining whether each episode is a planned admission, acute admission, or potentially planned admission (CMS, 2015). For each episode we then computed the waiting time to either the next unplanned acute episode or death, or until censorship due to the closure of the observation window. Altogether, the training dataset consisted of approximately 1.2 million inpatient episodes, of which approximately 17% were followed by an unplanned acute inpatient readmission and 3% were followed by death within 30 days. The histogram of the wait times is presented in Fig. 1.

4.1. Preprocessing and feature engineering

Medical claims data is expressed in several dialects (ICD9/10, HCPCS, RUG, HIPPS, etc). We converted codes for all procedures and diagnoses to a common clinically-curated dialect, the multilevel Clinical Classification Software (CCS) codes (AHRQ, 2022), maintained by the Agency for Health Research and Quality (AHRQ). The AHRQ also maintains databases that we used to tag comorbidities, chronic conditions, surgical flags, utilization flags, and procedure flags. From RUG and HIPPS codes we generated activities of daily living (ADL) scores, where higher scores correspond to lower functional ability. We incorporated CMS's risk adjustment methodology, hierarchical condition categories (HCC), into the model as covariates. For social determinants of health, we used race, medicaid state-buy in, the urban rural index, and social economic scale.

We used count encoding to turn codes into numerical feature vectors. In the case of CCS, which is multilevel, we aggregated and only counted to the first level (we also tried including second-order CCS coding but found it had only marginal impact on model performance). Altogether, the derived features constituted a vector of size $p \approx 7 \times 10^2$, including counts for the index episode and its four quarters of history at admission.

We placed all model parameters (log hazard ratios) on the same scale so that the magnitudes of all regression coefficients are directly comparable. In examining our derived data features, we found that they were predominantly sparse and heavy tailed. When fitting a logistic regression model to these data features, the model fit poorly to observations with large counts. Theses findings, and our desire to optimize model interpretability, led us to quantize all numerical variables so that the input variables into the model are entirely binary. To this end, we first computed the percentiles for each feature across the entire dataset. Then we re-coded each quantity into a series of binary variables corresponding to inequalities, where the cutoffs were determined by examining each variables at a set of quantiles and eliminating duplicate values. The usage of quantile-based coding has appeared in the literature (Hu et al., 2022; Saberian et al., 2019) as a nonlinear feature coding

that has demonstrated benefits to model performance in certain problems. Non-linearly transforming our count features using quantization improved the accuracy of logistic regression to nearly match that of XGBoost on this dataset as measured by AUROC. Hence, we used quantization for features in all models unless otherwise specified. The total size of the feature vector after dropping all original non-quantized numerical features and all constant features expanded to $p \approx 1 \times 10^3$.

5. Results

Using Medicare claims data, our intrinsically interpretable varying-coefficient survival model, and the guidance of Proposition 2, we studied the effect of postdischarge interventions in reducing readmission or death after hospital discharge.

5.1. Implementation

The `bayesianquilts`[1] library, built on top of Tensorflow Probability (Dillon et al., 2017), provides implementations of the piecewise exponential distribution, parameter decomposition, and approximate Bayesian inference scheme. We trained our model using minibatch mean-field stochastic ADVI, using minibatch sizes of 5×10^3, and a parameter sample size of 16 for approximating the variational loss function. We utilized the Adam optimizer with a starting learning rate of 0.0015. Each epoch where the mean batch loss did not decrease, we set the learning rate to decay by 10%. Training was set to conclude if there was no improvement for 3 epochs, or if we reached 50 epochs, whichever came sooner. We used scikit-learn 1.1.1 for fitting baseline logistic regression models, and XGBoost 1.6.1 for fitting a reference blackbox model for comparison. For logistic regression we provide two reference models: first a model restricted to only LACE predictors (Su et al., 2020) and second a model using all of our derived data features. We implemented a horseshoe Bayesian convolution neural network with ReLU activation using TFP, where we used a single hidden layer of size one-quarter the input layer. Similarly, we implemented exponential and Weibull Bayesian ReLU-nets using outputs of size one and two respectively. All computation was performed using the Pittsburgh Supercomputing Center's Bridges2 resources. We utilized extreme memory (EM) nodes for preprocessing, and Bridges2-GPU-AI for training.

Table 1 presents the classification accuracy of our model in predicting readmissions or death within the first 30 days, benchmarked against predictions given by alternative models trained on the same dataset. The standard deviation in both the AUROC and AUPRC measures, as determined using bootstrap, was approximately 0.003. The Bayesian neural network we developed utilizes sparsity-inducing horseshoe priors (Carvalho et al., 2010) on the weights and biases, which has been shown to improve model performance (Bhadra et al., 2019).

5.2. Readmission risk factors

We inferred a diverse set of pre-treatment baseline hazards (top of Fig. 3). The riskiest episode cohorts, and their definitions, are depicted in the bottom of Fig. 3. The baseline

1. `github:mederrata/bayesianquilts`

Model	Interpretability	AUROC/AUPRC	
30-day classification models		30-day	
XGBoost w/o quantization	None	0.741 / 0.465	
ReLU-BNN classifier	Computationally	0.750 / 0.481	
LR classifier - LACE only	Comprehensibly	0.666 / 0.313	
LR classifier	Comprehensibly	0.747 / 0.448	
Survival models		30-day	90-day
Exponential ReLU-BNN	Computationally	0.730 / 0.410	0.753 / 0.612
Exponential Cox PH	Comprehensibly	0.729 / 0.403	0.753 / 0.606
Exponential Quilt	Comprehensibly	0.744 / 0.468	0.760 / 0.632
Weibull AFT	Comprehensibly	0.533 / 0.207	0.530 / 0.362
Weibull ReLU-BNN	Comprehensibly	0.649 / 0.296	0.676 / 0.503
Weibull Quilt	Comprehensibly	0.688 / 0.334	0.717 / 0.551
Generalized Gamma Quilt	Comprehensibly	0.704 / 0.345	0.718 / 0.554
PEM quilt prediction-only	Comprehensibly	0.751 / 0.477	0.765 / 0.638
PEM quilt (post-discharge)	Mechanistically	0.753 / 0.468	0.774 / 0.641

Table 1: **30-day unplanned readmission or death classification metrics** for evaluated models: XGBoost, Sparse logistic regression (LR), Bayesian neural network (BNN) classifier, our quilted piece-wise exponential model (PEM), exponential Cox proportional hazards (PH), Weibull accelerated failure time (AFT), Weibull/exponential Bayesian neural networks, and quilted exponential, Weibull, generalized gamma models. Bootstrapped-estimated standard deviation for the AUROC ranged between 0.001 and 0.003 across the models. Quantization refers to the histogram-based bucketization of real-valued features. Area under the receiver operator curve (AUROC) and area under the precision-recall curve (AUPRC) computed on held-out 2012 inpatient episodes. Models trained on 2009-2011 episodes. Interpretability judged according to the criteria in Chang et al. (2022).

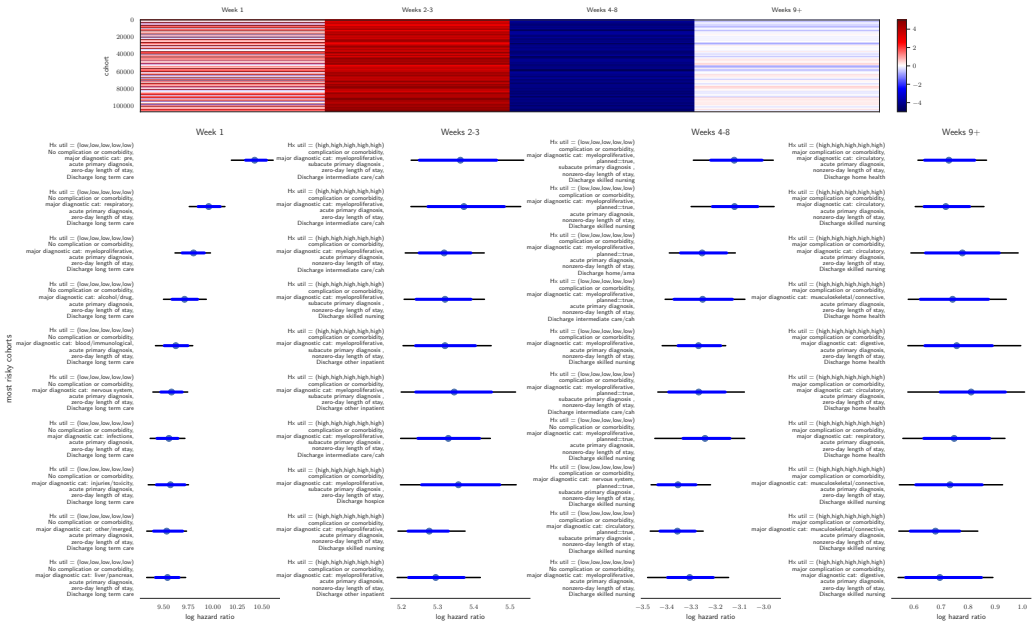

Figure 3: **Pre-treatment cohort-wise readmission/death risk** in each time interval, as defined by magnitude of log-hazard ratio. (top) Mean posterior log-hazard shown for all cohorts. (bottom) Mean, 80%, and 95% posterior credible intervals for the eight most-risky cohorts.

hazard cohorts are defined by interaction of beneficiary history grouping (see Supplemental Fig. 7), beneficiary age grouping, Medicaid state buyin, comorbid/complicated DRG, discharge location, and presence of acute primary diagnosis at index admission.

The most impactful contributors to variation in readmission/death hazards, as measured by the magnitude of the posterior mean effect size, are depicted in Fig. 4. We decomposed these model parameters so that they vary between twelve broad episode cohorts, defined by taking the interaction of three conditions. We allowed the regression coefficients to vary across these cohorts using the additive parameter decomposition method that we described.

The inferred treatment effects (posterior mean and standard deviation) are depicted in Fig. 5, for each episode cohort and time interval. The three interventions with the largest impact were office visit, nursing facility care, case management. For these three interventions, we also display the cohorts that exhibited the greatest first-week effects in Fig. 6.

6. Discussion

While not the focus of this manuscript, we demonstrate in Table 1 that we are able achieve accuracy comparable to that of the most popular black-box methods, without compromising model interpretability. As noted elsewhere (Rudin and Radin, 2019; Rudin, 2019; Chang et al., 2022; Ghassemi et al., 2021; Babic et al., 2021), posthoc explainer methods such as

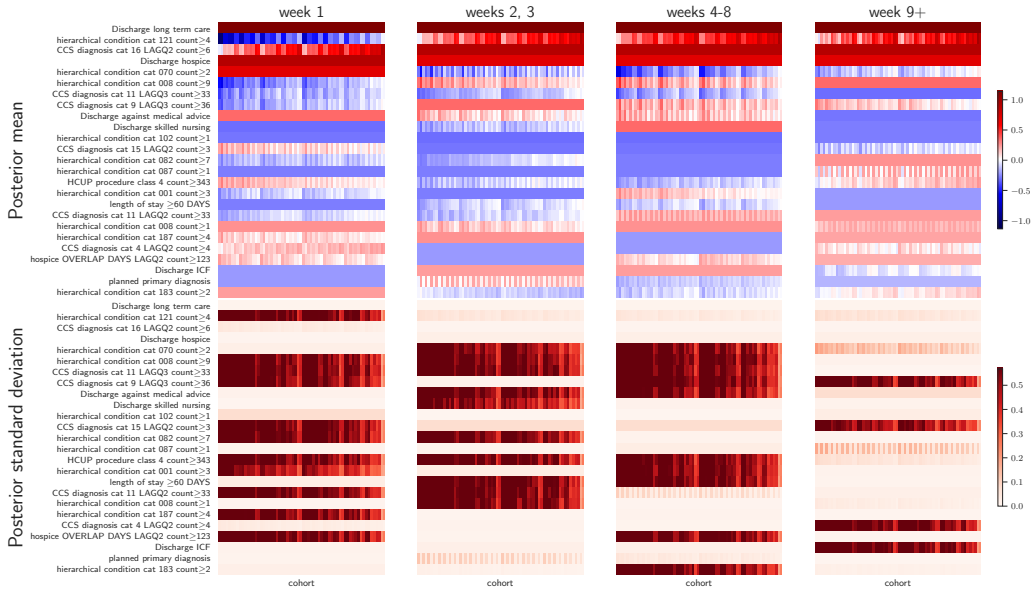

Figure 4: **Top predictors of readmission or death after discharge**, as defined by magnitude of log-hazard ratio, broken out by cohort. Cohorts delineated by interaction of conditions: Hx utlization, and whether the episode is expected to be followed by planned admissions.

SHAP are not reliable methods for making a blackbox understandable. Because our model is inherently interpretable (as a variant of regression), it easily admits **the** unequivocal model explanation – providing legitimate scientific insights, and enabling iterative human-guided criticism and improvement.

6.1. Model accuracy trends

In Table 1 we provided 30-day and 90-day classification metrics for several survival models, and 30-day classification metrics for common classification models. Looking at the classification models, logistic regression (LR) achieved performance comparable to XGBoost and deep learning (ReLU-BNN) as measured by the AUROC, though lagged behind both as measured by AUPRC. Logistic regression using only LACE predictors lagged behind all models in both measures.

The advantage of classification models (disadvantage of survival models) is that in using classification one can more-easily tune a model for discriminative performance at a given timepoint. Survival models are trained to fit the entire survival curve rather than focus on any single timepoint. The contrast between accuracy of these two types of models at 30 days illustrates this disadvantage. While the exponential, Weibull, and generalized gamma models are a nested family, performance did not reliably increase with increasing likelihood complexity. Examining the wait time histogram of Fig. 1, one sees that the overwhelming majority of episodes have wait time greater than 30-days – the Weibull distribution is more-robust than the exponential distribution in fitting the tail of this distribution, compro-

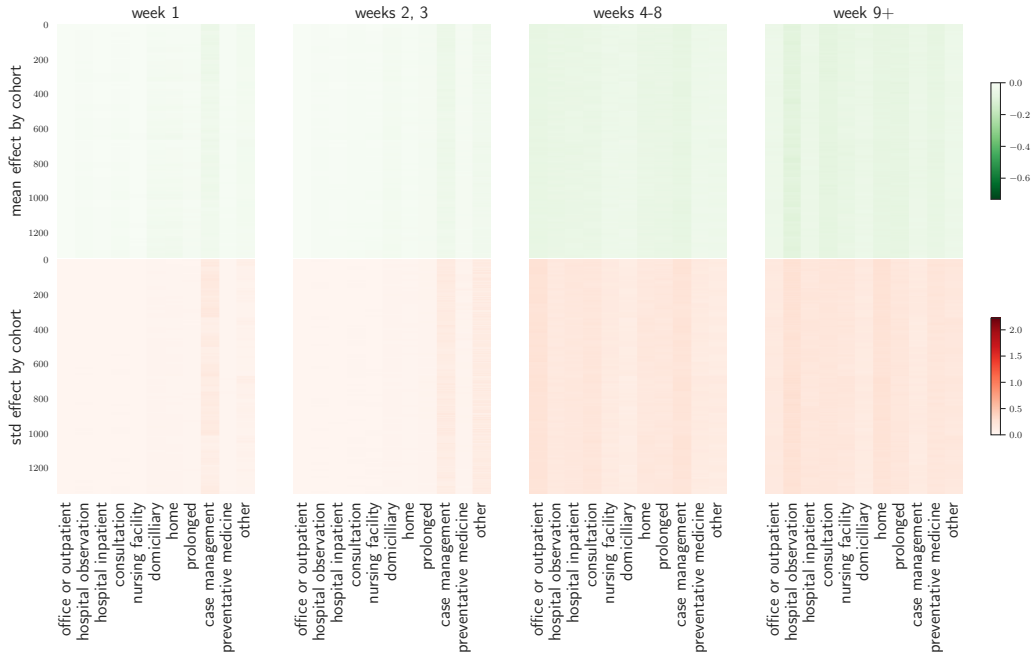

Figure 5: **Heterogeneous treatment effects for postdischarge evaluation management interventions.** Top is posterior mean and bottom is standard deviation. Effect varies across cohorts of like index hospital admissions.

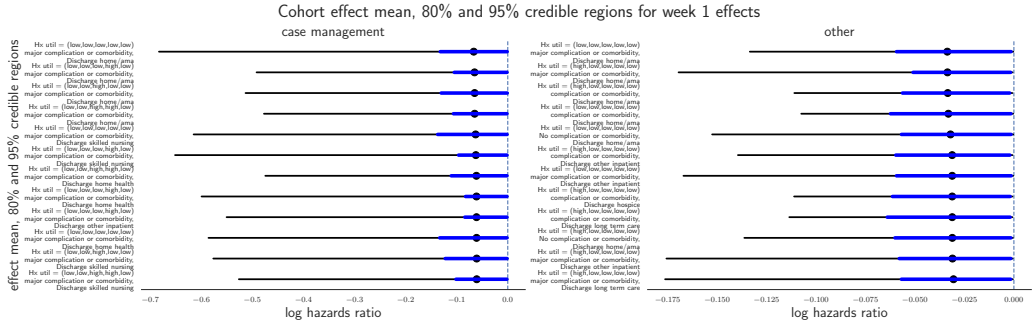

Figure 6: **Largest first-week treatment effects** for case management and E/M services categorized as "other."

mising performance at the early 30-day timepoint. Because we are able to explicitly specify timepoint breaks within piecewise exponential models, we were able to force the PEMs to better–focus on the early period after discharge. As a result, the PEMs had performance comparable to the top-performing classification models in predicting 30-day readmissions.

6.2. History utilization (Hx) episode grouping

We utilized the utilization representation model of Chang et al. (2022), based on sparsely-encoded Bayesian Poisson matrix factorization (Chang et al., 2021) to define factors used in cohort definition. Specifically, we used a five-dimensional embedding model and grouped each episode based on whether it exceeded the median value for each representation component (into low/high utilization per dimension). This procedure created $2^5 = 32$ disjoint history-based episode groups that were additionally interacted with other discrete grouping factors to define parameter decompositions. Each representation component is a linear combination of utilization variables (see Supplemental Fig. 7). Roughly, the first dimension corresponded most to CCS diagnosis categories 13 (diseases of the musculoskeletal system and connective tissue), 5 (mental illness), 16 (injury and poisoning), 4 (diseases of the blood and blood-forming organs), and CCS procedure categories 1 (operations on the nervous system), 14 (operations on the musculoskeletal system), across the entire preceding year of history. The second dimension corresponded most to CCS category 3 (endocrine; nutritional; and metabolic diseases and immunity disorders), 10 (diseases of the genitourinary system), 12 (diseases of the skin and subcutaneous tissue) diagnoses. The third dimension corresponded most to CCS category 7 (diseases of the circulatory system), 9 (diseases of the digestive system) diagnoses. The fourth dimension corresponded mostly to the recent quarter (past 90 days) of CCS category 7 (operations on the cardiovascular system), 9 (operations on the digestive system) procedures and CCS category 8 (diseases of the respiratory system), 9 diagnoses. The fifth dimension corresponded mostly to the number of outpatient encounters and CCS category 2 (neoplasms) diagnoses.

6.3. Risk predictors

In our intrinsically interpretable model, each model parameter is a log hazard ratio associated with a given predictor variable. In Fig. 3, the most-risky cohorts in Week 1 corresponded to those discharged the same day as admission, with low prior-year history utilization, and no complication or comorbidity as defined by DRG code. For Weeks 2 and 3, the most-risky cohorts were those who had high history utilization in the prior year, zero day length of stay, and a myeloproliferative DRG code. The absolute hazard ratios level off in Weeks 4-8 and increase somewhat from nine weeks onwards. In Fig. 4, many hierarchical condition categories make an appearance as strong predictors of readmission risk. Additionally, the discharge status and cuttoffs for counts of various diagnosis and procedure classes are strong predictors overall. However, as seen in this figure, the impact of these predictors varied across each of the 64 cohorts over which the parameters $\boldsymbol{\beta}$ are defined.

6.4. Postdischarge intervention effects

Using our model we inferred heterogeneous treatment effects for each of eleven categories of postdischarge interventions (Fig. 5). Overall, we found case management and E/M services categorized as "other" to have the greatest impact in terms of preventing readmission or death in the first four weeks after discharge, and beyond. In the first week the patient is the most vulnerable (Fig. 3). The cohorts that most-benefited from case management or "other" services were those episodes categorized by DRG as major complication or comorbidity, across all discharge codes and all history groupings.

6.5. Limitations

Our methodology requires input for model formulation. While aspects of the model structure are learned, most of the choices are intentional and based on a combination of domain knowledge, the desire to prioritize interpretability, and iterative model refinement. For our method, numerical stability generally requires the use of double precision floating point – the parameter decomposition is memory-intensive which in some applications may limit expressivity.

7. Conclusion

In this manuscript we introduced methodology for assessing treatment effects where the temporal nature of the outcome induces treatment selection bias. Failure to account for this bias leads to erroneous effect estimates – we derive the phantom effect associated with an ineffective intervention. Our approach uses survival regression, enhancing expressiveness by allowing regression coefficients to vary. We applied this methodology to the analysis of Medicare claims data, identifying specific cohorts of inpatient episode types that would best benefit from each category of postdischarge intervention. In particular, we found that case management services appear to have the largest impact in terms of reducing readmission risk.

8. Acknowledgments

We thank the Innovation Center of the Center for Medicare and Medicaid services for providing access to the CMS Limited Dataset through DUA LDSS-2019-54177. We also thank Dr. Pei-Shu Ho for help in understanding Medicare billing data. CCC is supported by the Intramural Research Program of the NIH, NIDDK. This work used the Extreme Science and Engineering Discovery Environment, which is supported by National Science Foundation grant number ACI-1548562 through allocation TG-DMS190042.

References

AHRQ. HCUP-US Tools & Software Page. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp, 2022.

Ahmed Allam, Mate Nagy, George Thoma, and Michael Krauthammer. Neural networks versus Logistic regression for 30 days all-cause readmission prediction. *Scientific Reports*, 9(1):9277, June 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-45685-z.

David Alvarez-Melis and Tommi S. Jaakkola. On the Robustness of Interpretability Methods. *arXiv:1806.08049 [cs, stat]*, June 2018.

Andrew Anderson, Carrie W. Mills, Jacqueline Willits, Craig Lisk, Jessica L. Maksut, Meagan T. Khau, and Sarah Hudson Scholle. Follow-up Post-discharge and Readmission Disparities Among Medicare Fee-for-Service Beneficiaries, 2018. *Journal of General Internal Medicine*, 37(12):3020–3028, September 2022. ISSN 1525-1497. doi: 10.1007/s11606-022-07488-3.

Rasha Assaf and Rashid Jayousi. 30-day Hospital Readmission Prediction using MIMIC Data. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6, October 2020. doi: 10.1109/AICT50176.2020.9368625.

Boris Babic, Sara Gerke, Theodoros Evgeniou, and I. Glenn Cohen. Beware explanations from AI in health care. *Science*, 373(6552):284–286, July 2021. doi: 10.1126/science.abg1834.

Joseph Bafumi and Andrew Gelman. Fitting Multilevel Models When Predictors and Group Effects Correlate, September 2007.

Anindya Bhadra, Jyotishka Datta, Yunfan Li, and Nicholas G. Polson. Horseshoe Regularization for Machine Learning in Complex and Deep Models. April 2019.

Damien Bricard and Zeynep Or. Does an Early Primary Care Follow-up after Discharge Reduce Readmissions for Heart Failure Patients? *Working Papers*, (DT73), March 2018.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, June 2010. ISSN 0006-3444. doi: 10.1093/biomet/asq017.

Joshua C. Chang, Patrick Fletcher, Jungmin Han, Ted L. Chang, Shashaank Vattikuti, Bart Desmet, Ayah Zirikly, and Carson C. Chow. Sparse encoding for more-interpretable feature-selecting representations in probabilistic matrix factorization. In *International Conference on Learning Representations*, 2021.

Joshua C. Chang, Ted L. Chang, Carson C. Chow, Rohit Mahajan, Sonya Mahajan, Joe Maisog, Shashaank Vattikuti, and Hongjing Xia. Interpretable (not just posthoc-explainable) medical claims modeling for discharge placement to prevent avoidable all-cause readmissions or death, August 2022.

CMS. 2015 Measure Information About the 30-Day All-Cause Hospital Readmission Measure, Calculated for the Value-Based Payment Modifier Program | Guidance Portal. https://www.hhs.gov/guidance/document/2015-measure-information-about-30-day-all-cause-hospital-readmission-measure-calculated, 2015.

Derek DeLia, Jian Tong, Dorothy Gaboda, and Lawrence P Casalino. Post-Discharge Follow-Up Visits and Hospital Utilization by Medicare Patients, 2007–2010. *Medicare & Medicaid Research Review*, 4(2):mmrr.004.02.a01, May 2014. ISSN 2159-0354. doi: 10.5600/mmrr.004.02.a01.

Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. TensorFlow Distributions. *arXiv:1711.10604 [cs, stat]*, November 2017.

Jianqing Fan and Wenyang Zhang. Statistical Methods with Varying Coefficient Models. *Statistics and its interface*, 1(1):179–195, 2008. ISSN 1938-7989.

Avi Feller, Todd Grindal, Luke Miratrix, and Lindsay C. Page. Compared to what? Variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics*, 10(3):1245–1285, September 2016. ISSN 1932-6157, 1941-7330. doi: 10.1214/16-AOAS910.

Michael Friedman. Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics*, 10(1):101–113, March 1982. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345693.

Joseph Futoma, Jonathan Morris, and Joseph Lucas. A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56:229–238, August 2015. ISSN 1532-0464. doi: 10.1016/j.jbi.2015.05.016.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, November 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00208-9.

Soumya Ghosh and Finale Doshi-Velez. Model Selection in Bayesian Neural Networks via Horseshoe Priors. *arXiv:1705.10388 [stat]*, May 2017.

Sara Nouri Golmaei and Xiao Luo. DeepNote-GNN: Predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21, pages 1–9, New York, NY, USA, August 2021. Association for Computing Machinery. ISBN 978-1-4503-8450-6. doi: 10.1145/3459930.3469547.

Kevin N. Griffith, David A. Schwartzman, Steven D. Pizer, Jacob Bor, Vijaya B. Kolachalama, Brian Jack, and Melissa M. Garrido. Local Supply Of Postdischarge Care Options Tied To Hospital Readmission Rates. *Health Affairs*, 41(7):1036–1044, July 2022. ISSN 0278-2715. doi: 10.1377/hlthaff.2021.01991.

Patricia L. Harrison, Pamela A. Hara, James E. Pope, Michelle C. Young, and Elizabeth Y. Rula. The Impact of Postdischarge Telephonic Follow-Up on Hospital Readmissions. *Population Health Management*, 14(1):27–32, February 2011. ISSN 1942-7891. doi: 10.1089/pop.2009.0076.

Trevor Hastie and Robert Tibshirani. Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993. ISSN 0035-9246.

Aaron Heuser, Minh Huynh, and Joshua C. Chang. Asymptotic convergence in distribution of the area bounded by prevalence-weighted Kaplan–Meier curves using empirical process modelling. *Royal Society Open Science*, 5(11):180496, November 2018. doi: 10.1098/rsos.180496.

Xinyu Hu, Tanmay Binaykiya, Eric Frank, and Olcay Cirit. DeeprETA: An ETA Post-processing System at Scale, June 2022.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, November 2020.

Xin Huang, Shande Chen, and Seng-jaw Soong. Piecewise Exponential Survival Trees with Time-Dependent Covariates. *Biometrics*, 54(4):1420–1433, 1998. ISSN 0006-341X. doi: 10.2307/2533668.

Yinan Huang, Ashna Talwar, Satabdi Chatterjee, and Rajender R. Aparasu. Application of machine learning in predicting hospital readmissions: A scoping review of the literature. *BMC Medical Research Methodology*, 21(1):96, May 2021. ISSN 1471-2288. doi: 10.1186/s12874-021-01284-z.

Mehdi Jamei, Aleksandr Nisnevich, Everett Wetchler, Sylvia Sudat, and Eric Liu. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PLoS ONE*, 12(7), July 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0181173.

Stephen F. Jencks, Mark V. Williams, and Eric A. Coleman. Rehospitalizations among patients in the Medicare fee-for-service program. *The New England Journal of Medicine*, 360(14):1418–1428, April 2009. ISSN 1533-4406. doi: 10.1056/NEJMsa0803563.

Rohan Khera and Harlan M. Krumholz. Effects of the Hospital Readmissions Reduction Program. *Circulation: Cardiovascular Quality and Outcomes*, 11(12):e005083, December 2018. doi: 10.1161/CIRCOUTCOMES.118.005083.

J.S. Kim and F. Proschan. Piecewise exponential estimator of the survivor function. *IEEE Transactions on Reliability*, 40(2):134–139, June 1991. ISSN 1558-1721. doi: 10.1109/24.87112.

John Kitchin, Naftali A. Langberg, and Frank Proschan. A NEW-METHOD FOR ESTIMATING LIFE DISTRIBUTIONS FROM INCOMPLETE DATA. *Statistics & Risk Modeling*, 1(3):241–256, March 1983. ISSN 2196-7040. doi: 10.1524/strm.1983.1.3.241.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. February 2022. doi: 10.48550/arXiv.2202.01602.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. *arXiv:2002.11097 [cs, stat]*, June 2020.

Chuhong Lahlou, Ancil Crayton, Caroline Trier, and Evan Willett. Explainable Health Risk Predictor with Transformer-based Medicare Claim Encoder, May 2021.

Anna Larsson, Johanna Berg, Mikael Gellerfors, and Martin Gerdin Wärnberg. The advanced machine learner XGBoost did not reduce prehospital trauma mistriage compared with logistic regression: A simulation study. *BMC Medical Informatics and Decision Making*, 21(1):192, June 2021. ISSN 1472-6947. doi: 10.1186/s12911-021-01558-y.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations, July 2019.

Feng Li, Yajie Li, and Sanying Feng. Estimation for Varying Coefficient Models with Hierarchical Structure. *Mathematics*, 9(2):132, January 2021. ISSN 2227-7390. doi: 10.3390/math9020132.

Wenshuo Liu, Cooper Stansbury, Karandeep Singh, Andrew M. Ryan, Devraj Sukul, Elham Mahmoudi, Akbar Waljee, Ji Zhu, and Brahmajee K. Nallamothu. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PLoS ONE*, 15(4), April 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0221606.

Xiong Liu, Yu Chen, Jay Bae, Hu Li, Joseph Johnston, and Todd Sanger. Predicting Heart Failure Readmission from Clinical Notes Using Deep Learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2642–2648, November 2019. doi: 10.1109/BIBM47256.2019.8983095.

Emily J. MacKay, Michael D. Stubna, Corey Chivers, Michael E. Draugelis, William J. Hanson, Nimesh D. Desai, and Peter W. Groeneveld. Application of machine learning approaches to administrative claims data to predict clinical outcomes in medical and surgical patient populations. *PLOS ONE*, 16(6):e0252585, June 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0252585.

Ganesh Malla and Hari Mukerjee. A new piecewise exponential estimator of a survival function. *Statistics & Probability Letters*, 80(23-24):1911–1917, 2010.

Colleen K. McIlvennan, Zubin J. Eapen, and Larry A. Allen. Hospital Readmissions Reduction Program. *Circulation*, 131(20):1796–1803, May 2015. ISSN 0009-7322. doi: 10.1161/CIRCULATIONAHA.114.010270.

Xu Min, Bin Yu, and Fei Wang. Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. *Scientific Reports*, 9(1):2362, February 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-39071-y.

Margaret Sullivan Pepe and Thomas R. Fleming. Weighted Kaplan-Meier Statistics: Large Sample and Optimality Considerations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):341–352, 1991. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1991.tb01827.x.

Nicholas G. Polson and James G. Scott. *Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction **. Oxford University Press, October 2011. ISBN 978-0-19-173192-1.

Stephen W. Raudenbush, Sean F. Reardon, and Takako Nomi. Statistical Analysis for Multisite Trials Using Instrumental Variables With Random Coefficients. *Journal of Research on Educational Effectiveness*, 5(3):303–332, July 2012. ISSN 1934-5747. doi: 10.1080/19345747.2012.689610.

William Rhodes. Heterogeneous Treatment Effects: What Does a Regression Estimate? *Evaluation Review*, 34(4):334–361, August 2010. ISSN 0193-841X. doi: 10.1177/0193841X10372890.

Donald B. Rubin. Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with "Censoring" Due to Death. *Statistical Science*, 21(3): 299–309, August 2006. ISSN 0883-4237, 2168-8745. doi: 10.1214/088342306000000114.

Cynthia Rudin. Algorithms for interpretable machine learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1519, New York, NY, USA, August 2014. Association for Computing Machinery. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2630823.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.

Cynthia Rudin and Joanna Radin. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, 1(2), November 2019. ISSN 2644-2353, 2688-8513. doi: 10.1162/99608f92.5a8a3a3d.

Mohammad Saberian, Pablo Delgado, and Yves Raimond. Gradient Boosted Decision Tree Neural Network, November 2019.

Khader Shameer, Kipp W Johnson, Alexandre Yahi, Riccardo Miotto, Li Li, Doran Ricks, Jebakumar Jebakaran, Patricia Kovatch, Partho P. Sengupta, Sengupta Gelijns, Alan Moskovitz, Bruce Darrow, David L David, Andrew Kasarskis, Nicholas P. Tatonetti, Sean Pinney, and Joel T Dudley. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using mount sinai heart failure cohort. In *Biocomputing 2017*, pages 276–287. WORLD SCIENTIFIC, November 2016. ISBN 978-981-320-780-6. doi: 10.1142/9789813207813_0027.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *arXiv:1911.02508 [cs, stat]*, February 2020.

Mei-Chin Su, Yi-Jen Wang, Tzeng-Ji Chen, Shiao-Hui Chiu, Hsiao-Ting Chang, Mei-Shu Huang, Li-Hui Hu, Chu-Chuan Li, Su-Ju Yang, Jau-Ching Wu, and Yu-Chun Chen. Assess the Performance and Cost-Effectiveness of LACE and HOSPITAL Re-Admission Prediction Models as a Risk Management Tool for Home Care Patients: An Evaluation Study of a Medical Center Affiliated Home Care Unit in Taiwan. *International Journal of Environmental Research and Public Health*, 17(3):927, January 2020. doi: 10.3390/ijerph17030927.

Agus Sudjianto and Aijun Zhang. Designing Inherently Interpretable Machine Learning Models, November 2021.

Sara van Erp, Daniel L. Oberski, and Joris Mulder. Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50, April 2019. ISSN 0022-2496. doi: 10.1016/j.jmp.2018.12.004.

Duncan Vernon, James E Brown, Eliza Griffiths, Alan M Nevill, and Martha Pinkney. Reducing readmission rates through a discharge follow-up service. *Future Healthcare Journal*, 6(2):114–117, June 2019. ISSN 2514-6645. doi: 10.7861/futurehosp.6-2-114.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do Feature Attribution Methods Correctly Attribute Features? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9623–9633, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i9.21196.

Appendix A. Proofs

Lemma 1 (The intervention-time-corrected time-to-event likelihood) *Let $T > 0$ denote the wait time to the next admission, or death, after discharge, where the statistics of the wait time depend on the time $\tau \geq 0$ that an intervention occurs, according to the conditional probability density function*

$$T|\tau \sim f_\tau(T). \tag{1}$$

where $f_\infty(T)$ is the wait time probability density function conditional on no intervention occurring. Suppose that the effect of an intervention is to modify the wait time statistics so that the post-intervention waiting time $(T - \tau)$ follows the density $g_\tau(T - \tau)$, where g_τ is allowed to depend on the time of intervention. Then, the effective probability density function of the total wait time is

$$f_\tau(T) = \begin{cases} f_\infty(T) & T < \tau \\ g_\tau(T-\tau)\left(1 - \int_0^\tau f_\infty(u)\mathrm{d}u\right) & T \geq \tau. \end{cases} \tag{2}$$

See proof in Appendix A.

Remark 2 *The effect of the intervention is null if and only if*

$$g_\tau(T-\tau) = \frac{f_\infty(T)}{\int_\tau^\infty f_\infty(u)\mathrm{d}u} \tag{3}$$

$\forall T > \tau$ because this relationship would imply that $f_\tau(T) = f_\infty(T)$, $\forall T \geq 0$.

Lemma 3 (Wait time distribution | observed) *Suppose as in Lemma 1 that the waiting time after intervention at time τ is distributed according to the density $g_\tau(T - \tau)$. If τ is distributed*

$$\tau \sim h(\tau), \tag{4}$$

*then, the wait time distribution for admissions **where no intervention is observed** is*

$$f(T|T < \tau) = f_\infty(T) \tag{5}$$

*and the wait time distribution **where the intervention is observed** is*

$$f(T|T \geq \tau) = \int_0^\infty g_\tau(T-\tau)\left(1 - \int_0^\tau f_\infty(u)\mathrm{d}u\right) h(\tau)\mathrm{d}\tau. \tag{6}$$

Proof By conditioning,

$$\begin{aligned} f(T|T \leq \tau) &= \int_0^\infty f(T|T \leq \tau, \tau) h(\tau)\mathrm{d}\tau \\ &= \int_0^\infty f_\infty(T) h(\tau)\mathrm{d}\tau \\ &= f_\infty(T). \end{aligned} \tag{7}$$

The argument for $f(T|T \geq \tau)$ is the same. ■

Theorem 1 (Phantom effect due to survivors bias) *Suppose that an intervention administered at time $\tau \sim h(\tau)$ has no effect, then the apparent cumulative readmission probability by time c for cases where interventions have been observed is*

$$\Pr(T \leq c | T \geq \tau) = \int_0^c \left(S_\infty(\tau) - S_\infty(c)\right) h(\tau) \mathrm{d}\tau, \tag{1}$$

where $S_\tau(t) = \int_t^\infty f_\tau(t)\mathrm{d}t$ is the survival function corresponding to the wait time distribution probability density function f_τ, until readmission conditional on receiving an intervention at time τ (f_∞ is the pre-treatment wait time distribution probability density function).

Proof [Proof of Theorem 1]Eq. 1 is found by substituting the expression in Remark 2 into the expression for the wait time distribution where no intervention is observed in Lemma 3 and integrating over the limits $\tau \in [0, c]$. ■

Proposition 4 (Multiple interventions) *Suppose that $0 < \tau_1 \leq \tau_2 \leq \ldots \leq \tau_N$ are fixed times for which N interventions are scheduled. Denote f_∞ the pre-treatment wait time probability density function, and $g_{\tau_1,\tau_2,\ldots\tau_K,\infty,\ldots}$ the probability density function for the remaining waiting time after the K–th intervention (occurring at time τ_K), if the $K+1$-st intervention is never applied. Then*

$$T | \tau_1, \tau_2, \ldots, \tau_N \sim f_{\tau_1,\tau_2,\ldots,\tau_N}(T) \tag{2}$$

where

$$f_{\tau_1,\tau_2,\ldots,\tau_N}(T) = \begin{cases} f_\infty(T) & T < \tau_1 \\ g_{\tau_1,\ldots,\tau_n,\infty,\ldots}(T-\tau_n) & \\ \quad \times \left(1 - \int_0^{\tau_1} f_\infty(T)\mathrm{d}T - \sum_{k=1}^{n} \int_{\tau_k}^{\tau_{k+1}} g_{\tau_1,\ldots,\tau_k,\infty}(T-\tau_k)\mathrm{d}T\right) & T \in (\tau_n, \tau_{n+1}]. \end{cases} \tag{3}$$

Proof

Consider a test function $\phi(T)$. As in Eq. 11 we decompose its expectation

$$\mathbb{E}(\phi(T)|\tau_1, \ldots, \tau_N) = \sum_{i=1}^{N+1} \mathbb{E}\left(\phi(T) | T \in [\tau_{i-1}, \tau_i)\right) \Pr\left(T \in [\tau_{i-1}, \tau_i)\right), \tag{8}$$

where we define $\tau_0 \equiv 0$ and $\tau_{N+1} \equiv +\infty$. Conditional on the $i-1$-st intervention having been applied, we have

$$\mathbb{E}\left(\phi(T) | T \in [\tau_{i-1}, \tau_i)\right) = \int_{\tau_{i-1}}^{\tau} \frac{g_{\tau_1,\tau_2,\ldots,\tau_{i-1},\infty}(T-\tau_{i-1})}{\int_{\tau_{i-1}}^{\tau_i} g_{\tau_1,\tau_2,\ldots,\tau_{i-1},\infty}(u-\tau_{i-1})\mathrm{d}u} \phi(T)\mathrm{d}T \tag{9}$$

$$\begin{aligned}\Pr\left(T \in [\tau_{i-1}, \tau_i)\right) &= \Pr\left(T < \tau_i | T \geq \tau_{i-1}\right) \Pr\left(T \geq \tau_{i-1}\right) \\ &= \int_{\tau_{k-1}}^{\tau_k} g_{\tau_1,\ldots,\tau_{k-1},\infty}(T - \tau_k)\mathrm{d}T \\ &\quad \times \left(1 - \int_0^{\tau_1} f_\infty(T)\mathrm{d}T - \sum_{k=1}^{i} \int_{\tau_k}^{\tau_{k+1}} g_{\tau_1,\ldots,\tau_k,\infty}(T - \tau_k)\mathrm{d}T\right) \qquad (10)\end{aligned}$$

■

Lemma 5 (The intervention-time-corrected time-to-event likelihood)
Let $T > 0$ denote the wait time to the next admission, or death, after discharge, where the statistics of the wait time depend on the time $\tau \geq 0$ that an intervention occurs, according to the conditional probability density function

$$T|\tau \sim f_\tau(T). \qquad (1)$$

where $f_\infty(T)$ is the wait time probability density function conditional on no intervention occurring. Suppose that the effect of an intervention is to modify the wait time statistics so that the post-intervention waiting time $(T - \tau)$ follows the density $g_\tau(T - \tau)$, where g_τ is allowed to depend on the time of intervention. Then, the effective probability density function of the total wait time is

$$f_\tau(T) = \begin{cases} f_\infty(T) & T < \tau \\ g_\tau(T - \tau)\left(1 - \int_0^\tau f_\infty(u)\mathrm{d}u\right) & T \geq \tau. \end{cases} \qquad (2)$$

Proof Consider the expectation of a test function ϕ applied to the wait time T. We construct this expectation by conditioning on T relative to τ

$$\mathbb{E}(\phi(T)|\tau) = \mathbb{E}(\phi(T)|\tau, T \geq \tau) \Pr(T \geq \tau|\tau) + \mathbb{E}(\phi(T)|\tau, T < \tau) \Pr(T < \tau|\tau), \qquad (11)$$

where the first term in Eq. 11 is the contribution to the expectation when the intervention is successfully performed and the second term is the contribution when the intervention is not performed. The expectation of ϕ, conditional on the event occurring before time τ is the expectation of the truncated distribution

$$\mathbb{E}(\phi(T)|\tau, T < \tau) = \frac{\int_0^\tau \phi(T) f_\infty(T)\mathrm{d}T}{\int_0^\tau f_\infty(T)\mathrm{d}T}. \qquad (12)$$

The probability of the condition $T < \tau$ is

$$\Pr(T < \tau|\tau) = \int_0^\tau f_\infty(T)\mathrm{d}T, \qquad (13)$$

so

$$\mathbb{E}(\phi(T)|\tau, T < \tau) \Pr(T < \tau|\tau) = \int_0^\infty \phi(T) f_\infty(T)\mathrm{d}T. \qquad (14)$$

In the case where $T \geq \tau$, we have by definition of g,

$$\mathbb{E}(\phi(T)|\tau, T \geq \tau) = \int_{\tau}^{\infty} \phi(T) g_{\tau}(T-\tau) \mathrm{d}T, \tag{15}$$

where the probability of the corresponding condition follows

$$\Pr(T \geq \tau|\tau) = \int_{\tau}^{\infty} f_{\infty}(T) \mathrm{d}T. \tag{16}$$

Substituting these Eqs. into Eq. 11 yields

$$\begin{aligned}\mathbb{E}(\phi(T)|\tau) &= \int_{0}^{\tau} \phi(T) f_{\infty}(T) \mathrm{d}T + \int_{\tau}^{\infty} \phi(T) g_{\tau}(T-\tau) \int_{\tau}^{\infty} f_{\infty}(u) \mathrm{d}u \mathrm{d}T \\ &= \int_{0}^{\infty} \phi(T) \left[(1 - H(T-\tau)) f_{\infty}(T) + H(T-\tau) g_{\tau}(T-\tau) \int_{\tau}^{\infty} f_{\infty}(u) \mathrm{d}u \right] \mathrm{d}T,\end{aligned} \tag{17}$$

where H is the unit step function. ■

Appendix B. EM Codes

HCPCS code ranges corresponding to the different evaluation and management code categories

```
office_or_outpatient = np.arange(99202, 99216).astype(str)
hospital_observation = np.arange(99217, 99227).astype(str)
hospital_inpatient = np.arange(99221, 99240).astype(str)
consultation = np.arange(99241, 99256).astype(str)
nursing_facility = np.arange(99304, 99319).astype(str)
domicilliary = np.arange(99324, 99338).astype(str)
domicilliary = np.concatenate([domicilliary, ["99339", "99340"]])
home = np.arange(99341, 99351).astype(str)
prolonged = np.arange(99354, 99417).astype(str)
case_management = np.arange(99366, 99369).astype(str)
care_plan = np.arange(99374, 99381).astype(str)
preventative_medicine = np.arange(99381, 99430).astype(str)
care_management = np.arange(99439, 99492).astype(str)
special_eval = np.arange(99450, 99459).astype(str)
newborn_care = np.arange(99460, 99464).astype(str)
cognitive = np.arange(99483, 99487).astype(str)
behavioral = np.array([99484]).astype(str)
psych = np.arange(99492, 99495).astype(str)
transitional = np.arange(99495, 99497).astype(str)
other = np.arange(99497, 99500).astype(str)
```

Appendix C. Supplementary Methods

C.1. Medicare data preprocessing

Here we describe some details on the choices we made in preprocessing that will help make our work reproducible. Kyle Barron's Medicare Documentation repository of Medicare data documenation is an excellent resource for acquainting oneself with this standardized dataset. Our first steps in processing the CMS LDS were to merge the files, originally organized by year, into long tables for each claim type. In the process, we renamed pre-2011 columns in the dataset to match 2011+ plus columns where-ever they differed. We will refer to the dataset using 2011 and beyond column names.

C.1.1. Episode Grouping

The CMS LDS consists of records organized into claims. Multiple claims can constitute a single period or episode of service. We determined episodes of the following types:

1. inpatient (inp)
2. skilled nursing facility (snf)
3. hospice (hosp)
4. outpatient (out, car)

For determining episodes, we grouped claims of each of the given types by person, and sorted by either the admission date (for inp, snf, hosp), or the claim through-date for (out, car).

Then for inp, snf, hosp, we merged successive claims into running episodes if they overlapped temporally, if the provider was the same and the intermediate discharge code indicates that the individual was not otherwise discharged home in between (we allow for distinct episodes with zero days of wait if a patient is discharged home and returns on the same day).

For out and car, we did the same merging with all claim types together, relaxing the need for the provider to match in an episode. Then we filtered for out/car episodes that did not overlap with inp, snf, hosp episodes – we determined these to be true outpatient episodes.

Then, for out and inp episodes, we determined if they corresponded to emergency department visits by looking for corresponding revenue center codes.

C.2. Model Specification

The specific decompositions that we used for each of the model terms are displayed in Table 2. The python package `bayesianquilts`[2], contains utilities for managing decompositions such as these.

We used a regularized horseshoe prior in order to encourage $\boldsymbol{\beta}$ to be sparse. Specifically, we applied a horseshoe prior to the leading order term in its parameter decomposition. After training a model consisting of only the leading order terms, we isolated the 60 predictors who had the absolute largest coefficients for further expansion.

2. github:mederrata/bayesianquilts

Parameter	Decomposition	Max order
$\boldsymbol{\alpha}$	MDC × Hx × CC/MCC × age × medicaid	3
$\boldsymbol{\beta}$	Hx × planned	2
$\boldsymbol{\gamma}$	MDC × Hx × CC/MCC × Acute Primary Dx × discharge	3
$\boldsymbol{\eta}$	MDC × Hx × CC/MCC × age × medicaid	2
$\boldsymbol{\nu}$	MDC × Hx × CC/MCC × age × medicaid	2
$\boldsymbol{\xi}$	MDC × Hx × CC/MCC × age × medicaid	2

Table 2: **Specific decompositions used per parameter to define cohorts**, where major diagnostic category (MDC) is of size 26, history (Hx) is of size 2^5, corresponding to low/high in each of the five dimensions, CC/MCC is of size 2

C.2.1. Training

We utilize TFP's ADVI routines, which utilize stochastic sampling in computation of the ELBO. For this reason, it is not uncommon for specific parameter combinations to be in highly improbable locations – which can trigger underflows. To avoid instabilities, re adjust the likelihood on a per-observation level, first computing the minimum finite value of the log likelihood and then setting any divergent values to the minimum finite value minus a fixed offset of 100. We use the soft-plus function as a default bijector for any parameters that are supposed to be non-negative.

Appendix D. Supplementary Results

Here are selected results omitted from the main text for space constraints. We will make additional model results available at `https://www.mederrata.org/readmission/`.

D.1. History representation

We utilized sparse probabilistic matrix factorization in order to obtain a low-dimension representation of personal medical history for the year prior to each episode. The encodings given by the model specify linear combinations of the original data features that define a representation of an episode's history. The representations then can be constituted into a predictive distribution for the original features by transformation against a decoding matrix (Fig. 8). Note that this method finds a subset of the input features that can be used to predict the value of all features.

D.1.1. Random slopes

Although we do not use this terminology in the main text, in the language of hierarchical mixed effects models the parameters $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ in the model are random slopes. In Fig. 9, we present the components of $\boldsymbol{\beta}$ of the largest magnitudes, across all time intervals.

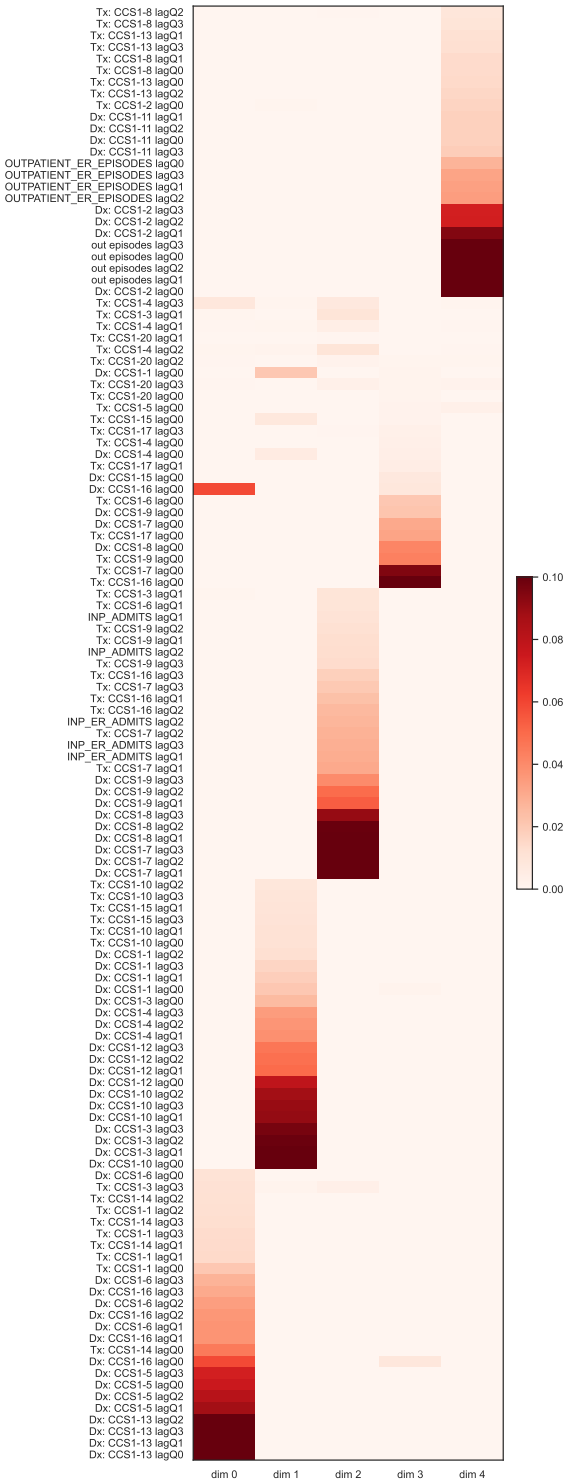

Figure 7: **Encoding matrix for utilization history model** with up to 25 features per dimension

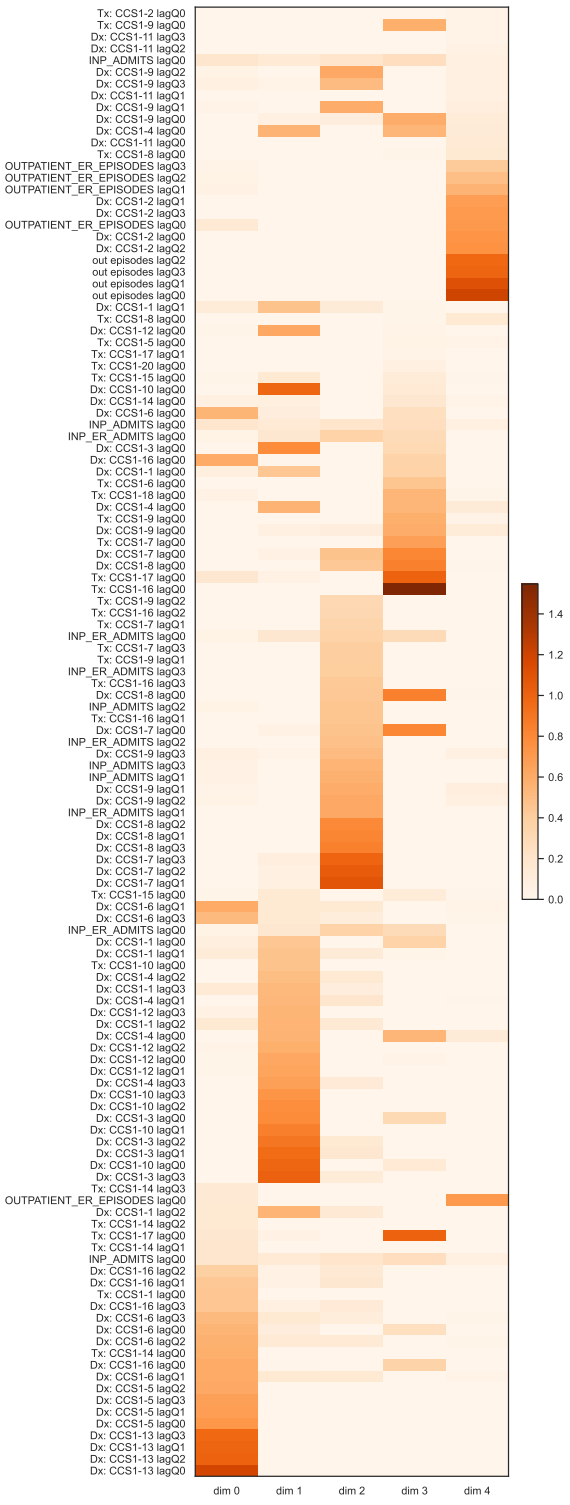

Figure 8: **Decoding matrix corresponding to the encoding model of Fig. 7 showing up 25 features per dimension**

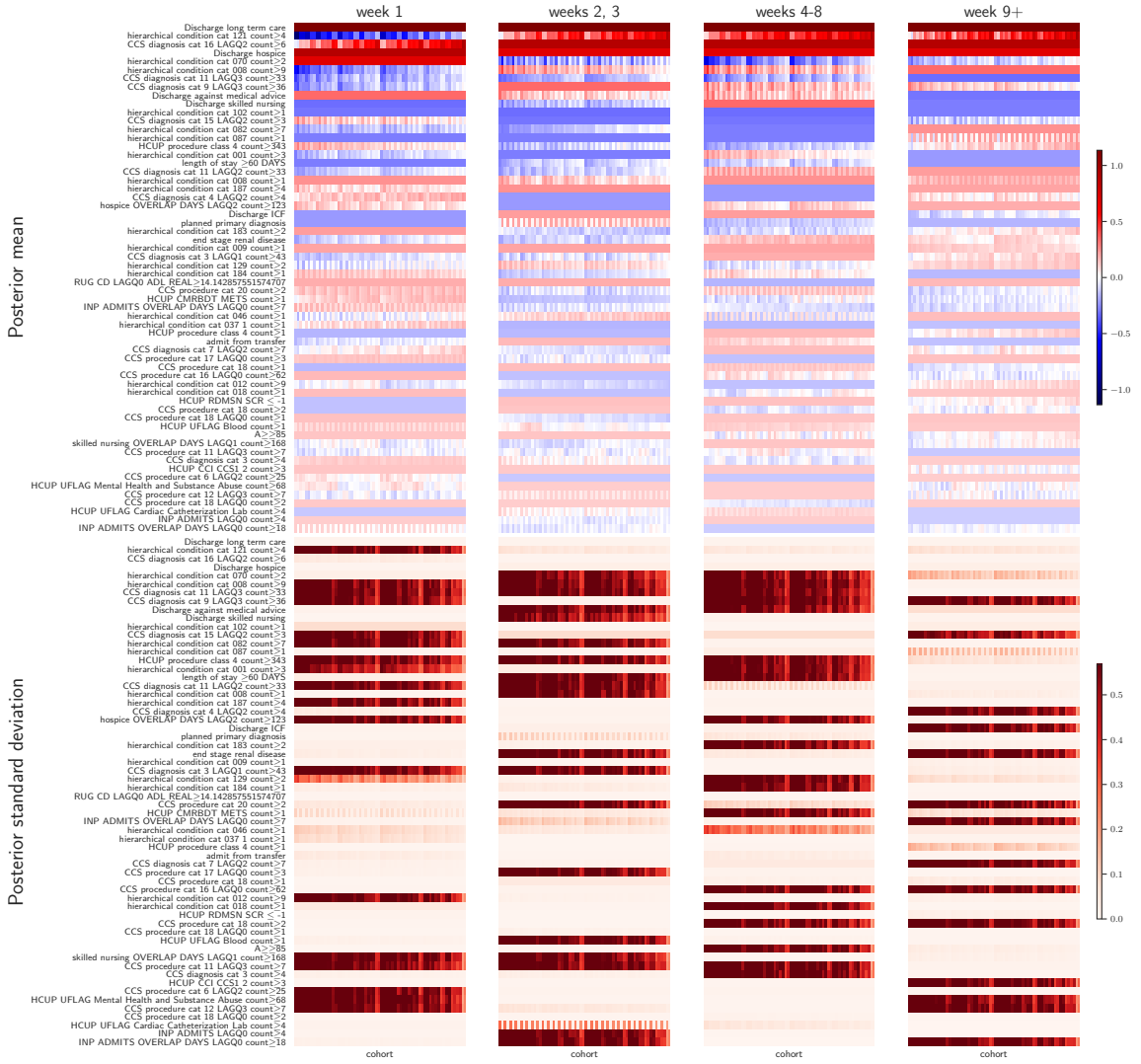

Figure 9: **The 50 most influential regressors β (posterior mean, standard deviation)** tracked through all time intervals. A more-comprehensive version of this figure can be found in our other supplemental file.