scientific reports

Check for updates

OPEN

Leveraging deep survival models to predict quality of care risk in diverse hospital readmissions

Nhat Quang Tran[1], Gautam Goel[1], Nirmala Pudota[2], Michael Suesserman[1], John Helms[1], Daniel Lasaga[3], Dan Olson[3], Edward Bowen[1] & Sanmitra Bhattacharya[1]✉

Hospital readmissions rate is reportedly high and has caused huge financial burden on health care systems in many countries. It is viewed as an important indicator of health care providers' quality of care. We examine the use of machine learning-based survival analysis to assess quality of care risk in hospital readmissions. This study applies various survival models to explore the risk of hospital readmissions given patient demographics and their respective hospital discharges extracted from a health care claims dataset. We explore advanced feature representation techniques such as BioBERT and Node2Vec to encode high-dimensional diagnosis code features. To our knowledge, this study is the first to apply deep-learning based survival-analysis models for predicting hospital readmission risk agnostic of specific medical conditions and a fixed window for readmission. We found that modeling the time from discharge date to readmission date as a Weibull distribution as in the SparseDeepWeiSurv model yields the best discriminative power and calibration. In addition, embedding representations of the diagnosis codes do not contribute to improvement in model performance. We find dependency of each model's performance on the time point at which it is evaluated. This time dependency of the models' performance on the health care claims data may necessitate a different choice of model in quality of care issue detection at different points in time. We show the effectiveness of deep-learning based survival-analysis models in estimating the quality of care risk in hospital readmissions.

Background. Hospital readmission rate is high. The rate of readmissions has been reported to be relatively high globally[1–4]. A study of hospital discharges of 12 million Medicare beneficiaries from 2 years of claims data reveals that nearly 20% of patients are readmitted within 30 days of discharge, 34% within 90 days, and over 56% within a year[1]. An analysis of 1306 inpatients aged 75 and older shows early unplanned readmissions happen at a rate of 14.2%[2]. Among patients with congestive heart failure, the readmission rate can be as high as 44% in 6 months[3]. This patient population is also among the highest early readmission rate in Canada and the United States (US)[4].

Hospital readmissions can place a huge financial burden on health care systems. In 2004, unplanned hospital readmissions accounted for USD 17.4 billion of the USD 102.6 billion paid by Medicare to hospitals[1]. In 2011, around 3.3 million adults in the US were readmitted within 30 days, associated with about USD 41.3 billion in hospital costs[5]. Canadian Institute for Health Information (CIHI) estimated a CAD 1.8 billion cost incurred by readmissions to acute care during an 11-month study period (excluding physician fees for services), accounting for 11% of total inpatient care costs[6].

Hospital readmission rate as an indicator of quality of care. In addition to incurring financial burdens on the health care system, hospital readmissions have also been viewed as red flags in hospitals' quality of care[7]. CIHI reports that between 9 and 59% of readmissions may be prevented by improving patient education, discharge planning, appropriately scheduling follow-up appointments, and conducting follow-up communications[8]. Reasons that may directly indicate quality of care, such as length of stay, have also been shown to have a direct contribution to hospital readmissions[1]. Boutwell and Hwu[9] suggests that for the patients with heart failure subgroup, the hospital readmission rate can be reduced by improved care, patient education, team management, and end-

[1]AI Center of Excellence, Deloitte & Touche LLP, New York, USA. [2]AI Center of Excellence, Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited, Hyderabad, India. [3]Program Integrity, Deloitte & Touche LLP, New York, USA. ✉email: sanmbhattacharya@deloitte.com

of-life care planning. In the US, the Centers for Medicare and Medicaid Services (CMS) established penalties for hospitals with high 30-day readmission rate by reducing the payment for readmitted patients[10]. In 2019, under the penalties program of CMS's Hospital Readmission Reduction Program, 82% of hospitals were penalized for having excess readmissions[11]. CMS includes the following six medical conditions to evaluate unplanned readmissions in the program:

- Acute myocardial infarction (AMI).
- Chronic obstructive pulmonary disease (COPD).
- Heart failure (HF).
- Pneumonia.
- Coronary artery bypass graft (CABG) surgery.
- Elective primary total hip arthroplasty and/or total knee arthroplasty (THA/TKA).

Besides the US, the United Kingdom (UK), Denmark, and Germany have also introduced policies, financial or non-financial, to monitor hospital readmission rates[12].

Objective. Since early hospital readmissions have been established as a measure to control for quality of care of medical services, our goal is to understand the risk of hospital readmissions given the information related to patients and their respective hospital discharges in Medicare/Medicaid claims data. Most previous studies have focused on the prediction of hospital readmission risk for comparisons among hospitals or for facilitating targeted interventions during or after hospital discharges[13]. These studies aim to predict the probability that a patient is readmitted within a specific time frame (usually 30 or 90 days), often using simple rule-based models such as the LACE index[14] or the HOSPITAL score[15]. A literature review by Ref.[16] reveals that 52 out of 76 studies use logistic regression to predict the likelihood of hospital readmission. Some other methods explored in prior research include support vector machines[17–21], decision tree-based techniques[17,22], Bayesian methods[22], and ensemble methods (e.g., boosting, bagging and random forest[4,17,18,20–24]). The majority of these studies structure the problem of hospital readmission risk prediction as a binary classification problem—whether a hospital discharge results in readmission within a certain number of days.

Another line of research is to learn a distribution of hospital readmission risk over time since an initial hospital discharge. For any time after discharge, these models predict the probability of hospital readmission occurring at or before the actual readmission time using survival analysis (or time-to-event analysis). The most commonly used survival model is the Cox Proportional Hazards model (Cox PH)[1,18,21,25–27]. A few studies have also implemented Random Survival Forest model, which decorrelates individual trees in the tree-based ensemble[27,28]. More recently, studies have shown that neural networks can improve the performance of traditional survival analysis models[29–35]. For example, DeepWeiSurv[30,31] uses a multi-task learning neural network on the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset (a UK-Canada project which tries to classify breast tumors into subcategories) and the Surveillance, Epidemiology, and End Results (SEER) dataset (which provides information on cancer statistics) to show that neural network based survival models outperform traditional survival models such as Cox PH. Similar to DeepWeiSurv, another fully parametric approach is Deep Survival Machines (DSM)[33]. DSM does not require constant proportional hazards assumption of the underlying survival distribution for time-to-event prediction. In contrast to DeepWeiSurv which learns the Weibull parameters and mixture coefficients from multi-layer perceptions following the latent representations, DSM learns these parameters directly from the latent representation.

Apart from restricting the analyses to a certain time frame after a hospital discharge, most studies focus on only one or a small set of medical conditions or diagnoses. Major conditions and diagnoses include heart failure[4,18,18,19,24–26,36], acute myocardial infarction[23,36,37], pneumonia[36,38], diabetes[22], and chronic obstructive pulmonary (COPD)[20].

Significance. To our knowledge, our study is the first to apply neural network-based survival-analysis models to predict hospital readmission risk from health care claims data, agnostic of specific medical conditions and a fixed window for readmission. There are several benefits to taking this approach in the context of quality of care. First, it allows us to identify quality of care issues for patients with *any* medical condition. This is especially important for claims data where patient populations are not segregated based on their diagnosis. Second, it gives us the probability of readmission within *any* time frame, making readmissions after the traditional 30-day or 90-day time frames also eligible for inspection on potential quality of care issues. While 30-day or 90-day time frames may be critical for policy compliance, these arbitrary time frames are not amenable to the diversity of medical conditions and corresponding discharge/ readmission times we consider in our study.

Materials and methods

A survival analysis framework is adopted in this study where a distribution over time to an event from a particular starting point is estimated. In our case, this *time-to-event* is the time elapsed between a hospital discharge and subsequent readmission for *similar* medical conditions. In survival analysis, typically, *censored* data needs to be handled. Censoring happens when a study subject is not being monitored or observed at a particular point in time (also known as censored time), and the occurrence of an event after the censored time is unknown. The two primary reasons for a data point to be censored in survival analysis are (1) a subject withdraws from the study so their information beyond the withdrawal time is unavailable, and (2) after a pre-specified cut-off time a subject is not monitored and hence survival data is not collected. In our study, censoring happens primarily due to the

latter reason as we do not consider hospital readmissions after $T = 1095$ days (3 years). While there may be other possible reasons for censoring, such as a patient changes their health insurance program and can no longer be tracked, or a patient expires at home, such events cannot be observed and hence not considered in our study.

Problem statement. In this section we formalize how we apply survival analysis to our data:

- A covariate matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ that represents d-dimension feature vectors of N hospital discharges. $x_n = \boldsymbol{X}[n][:]$ is the feature vector for the n-th discharge in the dataset.
- The time elapsed $t_n \in \mathbb{R}$ since the n-th discharge to either a readmission or a censored time.
- Censoring variable $\delta_n \in \{0, 1\}$ that indicates whether a readmission occurs at time t_n after the n-th discharge or it is censored at t_n.

Also denote T_n as the actual time of readmission following the n-th discharge ($T_n \equiv t_n$ if $\delta_n = 1$).
The goal is to estimate the distribution

$$f(t|x_n) \sim Pr(T_n = t|x_n, t_n, \delta). \tag{1}$$

Most survival models do not learn $f(t)$ directly. For example, the Cox PH model and its extensions (introduced in Models section) learn the hazard function:

$$\lambda(t) = -\frac{d}{dt}\log S(t), \tag{2}$$

where $S(t)$ is the survival function:

$$S(t) = \Pr(T_n \geq t). \tag{3}$$

The hazard function is the instantaneous rate of occurrence of the event at a particular time point and we can derive the desired density function from it.

Study data. We conducted this study on 222,175 redacted and anonymized inpatient medical claims from state Medicare programs. The dataset was redacted and anonymized following the Safe Harbor method, Section 164.514(b)(2) of the HIPAA Privacy Rule.

Data overview. A claim submitted to a Medicare program typically includes the following information:

- *Claim number* a distinct identifier of a claim.
- *Diagnosis codes* encoded using the International Classification of Diseases (ICD-10)[39], a standardized system used to encode clinical terms. A claim contains at least a primary diagnosis code, and optionally secondary and tertiary diagnosis codes. An ICD-10 code consists of up to 7 characters that distinctly identify a medical condition. The first three characters represent the general diagnosis, and the other characters represent more specific categories. Examples of a hierarchical break-down for general diagnosis codes I05 and I06 are shown in Table 1.

  When considering readmissions, we only analyze the first three digits of the ICD-10 codes, which represent the general category of the diagnoses. This helps us generalize our model by considering related diagnosis for which a patient may be readmitted.
- *Procedure codes* represented by Current Procedural Terminology (CPT) codes[40], encodes procedures performed by health care providers.
- *Provider ID* a unique identifier that represents the health care provider that submitted the claim, as used in National Provider Identifier (NPI)[41] registry.
- *Patient demographics* patient sex and age.
- *Admittance date and discharge date* the dates when a patient is admitted and discharged.
- *Billed amount* the total amount billed for services rendered by the health care provider.

Code	Description
I05-I09	Chronic rheumatic heart diseases
I05	Rheumatic mitral valve diseases
I05.0	Rheumatic mitral stenosis
I05.1	Rheumatic mitral insufficiency
I06	Rheumatic aortic valve diseases
I06.0	Rheumatic aortic stenosis
I06.1	Rheumatic aortic insufficiency

Table 1. Hierarchical breakdown of diagnosis codes.

Data pre-processing and representation. From a dataset of over 8 million claims, we filter for only inpatient claims, where a patient gets admitted to a hospital, and claims that have a positive paid amount. We construct two subsets: *readmission claims* and *non-readmission claims*. The readmission subset includes initial admissions and the subsequent readmissions of the same patient with the same general diagnosis codes. The non-readmission subset includes admissions without any subsequent readmissions.

To represent the dataset in a way that conforms to the structure of the survival data, we define a *time-to-event* and *censoring indicator* for each admission and readmission. Figure 1 illustrates how the data is represented. The time-to-event of each admission is the time (in days) from the discharge date of that admission to the admittance date of the next readmission. For admissions that are not followed by a readmission, the exact time-to-event is unknown, so we use the time from the discharge date to the last recorded date in the data (01/26/2020). These admissions are said to be *censored*.

Based on patient ID numbers, we split 222,175 total claims into training, validation, and test sets. Because of the severe class imbalance (readmission cases comprise of only 13% of the entire dataset), we downsample the training subset by first sampling one claim per patient in the non-readmitted set and then subsampling from these claims so that the size of the non-readmitted and that of the readmitted datasets are equal. Table 2 shows the observed vs. censored ratio in the training, validation, and test datasets. Supplementary Appendix A shows the dimension of each feature and the corresponding summary statistics (for applicable variables) for $n = 222,175$ claims.

Feature engineering. We use the following features in our survival analysis.

- *Patient age* Age bucketed into five categories: 0–17 years, 18–38 years, 39–59 years, 60–80 years, and greater than 80 years. We one-hot encode this feature.
- *Patient sex* patient sex is binarized into 0 (female) and 1 (male). No non-binary sex is present in the data.
- *Specialty code* this code represents the specialty[42] of the respective health care provider and is one-hot encoded. Empty code is represented as the 'UNK' (unknown) category.
- *Length of hospital stay* the difference in days between the discharge date and the admittance date. Claims with zero length of stay along with those within the same readmission chain with these claims are removed.
- *Diagnosis code* each claim number has at least one and at most three diagnosis codes. The primary code is always present. We only consider the first three digits of the codes as the general category of the diagnosis. For each claim number (a data point), we collect all the general diagnosis codes and multi-hot encode them. We do not consider codes that appear less than 100 times in the entire dataset and code them into an *Other* category.

Claim Number	Primary Diagnosis Code	Patient ID	Admittance date	Discharge date
A0001	F25	P0123	10/04/19	10/16/19
A0002	F25	P0123	11/01/19	11/11/19
A0003	F25	P0123	01/10/20	01/13/20
B0001	J01	T4920	12/29/19	01/21/20

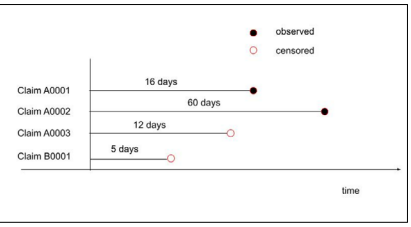

Figure 1. An example of data for four claims and how they are represented for survival analysis. *Left* Patient P0123 has three admissions (associated with three claim numbers) for the same diagnosis coded as F25 (schizoaffective disorders) and Patient T4920 only has one admission. *Right* For each claim number, the *time-to-event* is the time from the discharge date to the admittance date of the subsequent admission. For example, Claim A0001's time-to-event is 16 days (10/16/19–11/01/19). If an admission does not have any subsequent readmission (A0003 and B0001 in this case), the *time-to-event* is the time from the discharge date to the latest date recorded in the dataset 01/26/2020, and its censoring indicator is marked as *censored*. For example, B0001's time-to-event is 5 days (01/21/20–1/26/20) and is marked as censored. This is interpreted as the time-to-event for this discharge is *at least* 5 days, but we do not know when exactly the readmission will happen after 5 days (could be indefinitely long). A0001–0003 claims are in the readmission subset. B0001 is in the non-readmission subset. All claims in the non-readmission subset are censored. The last claim in each readmission chain in the readmission subset (e.g., claim A0003) is censored. The other claims are *observed*.

	# Observed/total (%)
Training	27,769/91,163 (30.46%)
Validation	6914/22,805 (30.31%)
Test	8747/108,207 (8.08%)

Table 2. Prevalence of observed cases in datasets.

Besides the multi-hot encoding approaches for feature representation, we experimented with word embedding and graph embedding models for feature representation of the diagnosis codes. For word embedding, we use the embeddings (dimension: 200) obtained from the BioWordVec model[43] trained on the string descriptions of the diagnosis codes. For graph embedding (dimension: 256), we use the Node2Vec model[44] trained on five million claims from the same dataset in this study, with the proxy task of link prediction (predicting whether or not a link exists between a pair of nodes).

While other features from the claims data could be viewed as relevant to readmission modeling, similar to previous studies[45,46], we focus on patient demographics and diagnosis as key indicators in estimating the likelihood of readmission. All the features listed above are concatenated into a vector representing features for the respective claim.

Models. We conduct a series of experiments to evaluate five survival analysis models (detailed in the next section) on our data.

Cox proportional hazards (Cox PH). Cox PH is one of the most common regression models and baseline models in survival analysis[1,21,25–27,47]. Cox PH assumes linearity to model the hazard function:

$$\lambda(t|x_n) = \lambda_0(t)e^{\beta \cdot x_n}. \tag{4}$$

Notice that $\lambda_0(t)$, the *baseline hazard*, depends on time but does *not* depend on the covariates x_n. It describes the risk of readmission when $x_n = 0$, and the exact risk level of each discharge is scaled by the exponential that depends on x_n (*proportional hazard assumption*). For this reason, the value $exp\{\beta \cdot x_n\}$ (also known as the *hazard ratio*) is characteristic of an individual's relative risk level compared to other individuals. Cox PH can be viewed as a regression model, which tries to estimate β to maximize the partial likelihood of data:

$$L(\beta) = \prod_{n:\delta_n=1} L_n(\beta), \tag{5}$$

where

$$L_n(\beta) = \frac{\lambda(t_n|x_n)}{\sum_{m:t_m \geq t_n} \lambda(t_n|x_m)}. \tag{6}$$

C-mix model. Besides Cox PH, we also include the C-mix model in our experiment, which is reported to outperform other survival models experimented in Ref.[21], a study on predictive models for hospital readmissions following vaso-occlusive crisis (VOC). C-Mix was originally designed to identify subgroups in the data with varying risk levels. It models the density of time-to-event as a mixture of Weibull distributions. In the case of a two-component mixture, the density is:

$$f(t|x_n) = \pi_0(x_n; \beta_0)f_0(t; \alpha_0) + \pi_1(x_n; \beta_1)f_1(t; \alpha_1), \tag{7}$$

where

$$\pi(x; \beta_k) = \frac{e^{x \cdot \beta_k}}{\sum_{k=0}^{1} e^{x \cdot \beta_k}},$$

where $\pi_0(\cdot)$ and $\pi_1(\cdot)$ are the weights for the components, f_1 and f_2 are two Weibull distributions parameterized by vectors α_1 and α_2, respectively, and:

$$\pi_i(x; \beta_i) = \exp(x \cdot \beta_i) / \left(\sum_{j \in \{1,2\}} \exp\left(x \cdot \beta_j\right) \right), i \in 1, 2. \tag{8}$$

The two components f_0 and f_1 in the mixture can be viewed as representing two subgroups: the high risk and the low risk groups, respectively. Then, $\pi_0(x_n; \beta_0)$ can be interpreted as the risk level. The parameters are estimated by minimizing the negative log likelihood of data.

DeepSurv. DeepSurv[29] is an extension of Cox PH as it uses a neural network to model the hazard function:

$$\lambda(t|x_n) = \lambda_0(t)e^{h_\theta(x_n)}, \tag{9}$$

where $h_\theta(\cdot)$ is a multilayer perceptron (MLP) with weights θ. Notice that this is almost identical to Eq. (1) except that the relationship with the covariates is modeled by an MLP instead of a linear function. DeepSurv is optimized by minimizing the negative log partial likelihood with regularization:

$$L_{\text{DeepSurv}} = -\frac{1}{N_{\lambda_n=1}} \sum_{n:\delta_n=1} \left(h_\theta(x_n) - log \sum_{j:t_j \geq t_i} e^{h_\theta(x_j)} \right) + \lambda \|\theta\|_2^2, \tag{10}$$

where λ is the regularization strength.

Sparse DeepWeiSurv. DeepWeiSurv[30] models the density over time to readmission as a mixture f_W of K Weibull distributions:

$$f(t|x_n, \theta_n) = \sum_{k=1}^{K} \alpha_k(x_n) f_{\beta_k(x_n), \eta_k(x_n)}, \tag{11}$$

where α_k is the weight, and β_k and η_k are the shape and scale parameters of the k-th Weibull component in the mixture (note that these parameters depend on the covariates x_n). The goal of DeepWeiSurv is to learn $\alpha \in \mathbb{R}^K, \beta \in \mathbb{R}^K, \eta \in \mathbb{R}^K$ from each x_n. DeepWeiSurv adopts a multi-task learning approach: there is a common sublayer f_{DWS} that is a MLP that learns a representation of x_n:

$$z_n = f_{DWS}(x_n). \tag{12}$$

After that DeepWeiSurv learns two MLP's f_1 and f_2:

$$\alpha(x_n) = f_1(z_n), \tag{13}$$

$$\beta(x_n), \eta(x_n) = f_2(z_n). \tag{14}$$

SparseDeepWeiSurv[31] extends DeepWeiSurv by incorporating a sparsing layer in f_1 to learn the number of components in the mixture. The model is optimized by minimizing the negative log likelihood. SpraseDeepWeiSurv outperforms DeepWeiSurv across five real-world datasets.

Deep cox mixture (DCM). DCM[34] fuses the Cox PH and DeepSurv to obtain a deep learning model that learns a mixture of Cox PH to model individual time-to-event distribution. It assumes there are latent groups and within each group, the proportional hazard assumption holds. In each Cox group of the mixture, DCM fits the hazard ratios using deep neural networks and the baseline hazard for each mixture component non-parametrically. It is reported to have a state-of-the-art performance on time-to-event regression tasks on survival data on mortality (e.g., METABRIC, SEER[30]).

Evaluation metrics. As pointed out by Ref.[34], most studies on survival models evaluate them using the relative ranking of the predictions of the risk level such as the concordance index (C-index). However, these metrics disregard the absolute values of the probability predictions, while these probabilities are directly used when detecting quality of care issues. The set of metrics that only depend on the ranking in terms of risk level of data points measure models' *discriminative power,* while those factoring in the actual predicted probabilities of readmission measure *calibration.* Following[34], we assess the statistical models on both aspects with the following 4 metrics. All metrics are time-dependent. In this study, we evaluate the metrics at the time points that are the 25th, 50th and 75th percentile of event times in our dataset.

Time-dependent concordance index (discrimination metric). The C-index measures the proportion of all eligible pairs of observations that are correctly ranked in terms of risk. The time-dependent concordance index restricts these comparisons to instances that occur within a certain time frame.

$$C(\boldsymbol{t}_0) = \frac{\sum_{i,j} \delta_i 1(\boldsymbol{t}_i < \boldsymbol{t}_j) 1(\boldsymbol{t}_i \leq \boldsymbol{t}_0) 1(\boldsymbol{\lambda}(\boldsymbol{t}_i|\boldsymbol{x}_i) > \boldsymbol{\lambda}(\boldsymbol{t}_j|\boldsymbol{x}_j))}{\sum_{i,j} \delta_i 1(\boldsymbol{t}_i < \boldsymbol{t}_j) 1(\boldsymbol{t}_i \leq \boldsymbol{t}_0)}. \tag{15}$$

Area under the receiver operation characteristic curve (AUC) (discrimination metric). At any point in time t_0, we can retrieve a binary label for any data point that indicates whether the readmission has happened by that time. Using $\lambda(t_0|x_i)$ to score an example i, we can compute the AUC as in a typical binary classification problem (using logistic regression for example).

Expected calibration error (ECE) (calibration metric). ECE is the average absolute difference between the observed and the predicted readmission rate, given the predicted readmission rate. Let the predicted readmission rate at time t_0 be $R(t_0|x_i) = \widehat{P}(t_i < t_0|x_i)$, then

$$ECE(t_0) = E(|R(t_0) - P(T > t_0|R(t_0))|). \tag{16}$$

We can estimate $ECE(t_0)$ by bucketing $R(t_0)$.

Brier score (dual metric). Brier score computes the mean squared error that quantifies the deviation of the predicted readmission rate within a time frame from the censoring indicator.

$$BR(t_0) = \frac{\sum_i 1(t_i > t_0)(0 - R(t_0|x_i))^2 + 1(t_i \leq t_0)\delta_i(1 - R(t_0|x_i))^2}{\sum_i 1(t_i > t_0) + 1(t_i \leq t_0)\delta_i}. \tag{17}$$

For model tuning and validation, we use a vanilla (non-time-dependent) version of the concordance index, which is traditionally used to evaluate survival analysis models and computed as:

$$C = \frac{\sum_{i,j} 1(t_j < t_i) 1(\lambda_j > \lambda_i) \delta_j}{\sum_{i,j} 1(t_j < t_i) \delta_j}. \tag{18}$$

Results

Experiment setting. For deep models (DeepSurv, SparseDeepWeiSurv, and DCM), we tune the models' hyper-parameters based on the computed concordance index on the validation subset. Further training details are provided in Supplementary Appendix B.

Results. For each of the five models, following the approach in Ref.[34], we compute the four evaluation metrics at three time-quantiles, 25th, 50th and 75th ones. The 25th, 50th and 75th time-quantiles correspond to readmission time frames of 17 days, 49 days, and 123 days. The metrics are reported in Table 3 for the test dataset.

For a short window (e.g., 25th percentile), the DCM model has the highest discriminative power, with AUC of 0.822 and C-index of 0.817, closely matched by SparseDeepWeiSurv, with AUC of 0.821 and C-index of 0.815. SparseDeepWeiSurv, on the other hand, is the best calibrated with the lowest Brier score and ECE. Notably, its ECE at 0.007 is 79% lower than the second lowest ECE of 0.034 achieved by DCM.

For larger time windows (i.e., 50th and 75th percentiles), SparseDeepWeiSurv outperforms other models in both calibration and discrimination with best performance across all four metrics. For these larger time windows, Cox PH closely matches SparseDeepWeiSurv on discrimination metrics (e.g., for 50th percentile window, both Cox PH and SparseDeepWeiSurv have a C-index of 0.827). With respect to ECE, as with a smaller percentile window, the gap between the performance of SparseDeepWeiSurv is large (80% and 69% lower than the second-lowest ECE in 50th percentile and 75th percentile windows, respectively). DCM has lower discriminative power but better calibration compared to Cox PH. C-mix and DeepSurv models consistently have the lowest performance across almost all metrics and percentiles, with the exception of C-mix's ECE of 0.060 at 50th percentile, where it achieves the second best calibration score.

As discussed in "Feature engineering" section, we also experimented with word and graph embedding based feature representation of the diagnosis codes. The results of using these embeddings are reported in Tables 4 and 5.

Using word embeddings of the diagnosis codes, we observe a minor improvement in the best calibration score ECE. For example, the best ECE at 25th percentile when using multi-hot encoded diagnosis codes is by SparseDeepWeiSurv (0.007 ECE), and the corresponding figure for word embedding is 0.004. However, with metrics at 75th percentile, the ECE performs worse than in the multi-hot encoding experiments (e.g., for SparseDeepWeiSurv, performance is worsened from 0.040 to 0.043). Using graph embeddings, we observe a decrease in performance across all metrics and models. Overall, we do not see significant improvement in model performance when incorporating advanced embedding techniques for embedding health care diagnosis codes.

Finally, we conduct an experiment with a 30-day readmission time frame. This time frame is commonly used in the existing literature on hospital readmission analysis and makes our finding comparable to other studies. Since word or graph embeddings of diagnosis codes do not improve model performance, we conduct this experiment with multi-hot encoding of diagnosis codes. Results in Table 6 show that the DCM model has the highest discriminative power with AUC and C-index scores of 0.836 and 0.831, respectively. SparseDeepWeiSurv is the best calibrated model with the lowest Brier score and ECE of 0.029 and 0.009, respectively.

	AUC	C-index	Brier score	ECE
25th percentile				
Cox PH	0.817	0.812	0.021	0.037
C-mix	0.801	0.795	0.021	0.020
DeepSurv	0.801	0.795	0.022	0.037
DCM	**0.822**	**0.817**	0.022	0.034
SparseDeepWeiSurv	0.821	0.815	**0.020**	**0.007**
50th percentile				
Cox PH	0.832	**0.827**	0.043	0.076
C-mix	0.817	0.809	0.041	0.060
DeepSurv	0.814	0.807	0.045	0.077
DCM	0.832	0.825	0.043	0.071
SparseDeepWeiSurv	**0.834**	**0.827**	**0.036**	**0.012**
75th percentile				
Cox PH	0.839	0.827	0.067	0.118
C-mix	0.826	0.814	0.073	0.121
DeepSurv	0.819	0.808	0.072	0.121
DCM	0.829	0.818	0.072	0.114
SparseDeepWeiSurv	**0.841**	**0.829**	**0.053**	**0.040**

Table 3. AUC, C-index, Brier score, and ECE computed at the 25th, 50th and 75th percentiles for the five models on the test set. The diagnosis codes are multi-hot encoded. The best performing value for each evaluation metric is highlighted in bold.

	AUC	C-index	Brier score	ECE
25th percentile				
Cox PH	0.800	0.795	0.021	0.037
C-mix	0.778	0.774	0.020	0.022
DeepSurv	0.801	0.795	0.022	0.035
DCM	0.801	0.797	0.021	0.034
SparseDeepWeiSurv	**0.805**	**0.799**	**0.019**	**0.004**
50th percentile				
Cox PH	0.812	0.805	0.043	0.076
C-mix	0.787	0.781	0.042	0.066
DeepSurv	0.813	0.806	0.045	0.035
DCM	0.810	0.804	0.042	0.072
SparseDeepWeiSurv	**0.818**	**0.811**	**0.035**	**0.008**
75th percentile				
Cox PH	0.820	0.809	0.069	0.117
C-mix	0.797	0.787	0.075	0.128
DeepSurv	0.818	0.807	0.074	0.116
DCM	0.800	0.789	0.073	0.124
SparseDeepWeiSurv	**0.824**	**0.813**	**0.053**	**0.043**

Table 4. AUC, C-index, Brier score, and ECE computed at the 25th, 50th and 75th percentiles for the five models on the test set. The diagnosis codes are embedded using word embedding. The best performing value for each evaluation metric is highlighted in bold.

	AUC	C-index	Brier score	ECE
25th percentile				
Cox PH	0.809	0.805	0.021	0.036
C-mix	0.796	0.793	**0.020**	0.022
DeepSurv	0.813	0.807	0.021	0.035
DCM	**0.818**	**0.814**	0.021	0.036
SparseDeepWeiSurv	0.810	0.804	**0.020**	**0.019**
50th percentile				
Cox PH	0.823	0.815	0.043	0.074
C-mi	0.806	0.799	0.042	0.064
DeepSurv	0.827	0.820	0.042	0.074
DCM	**0.828**	**0.821**	0.043	0.075
SparseDeepWeiSurv	0.823	0.816	**0.040**	**0.052**
75th percentile				
Cox PH	0.831	0.819	0.069	0.115
C-mix	0.812	0.801	0.074	0.124
DeepSurv	**0.836**	**0.824**	0.067	0.115
DCM	0.821	0.810	0.073	0.120
SparseDeepWeiSurv	0.828	0.817	**0.066**	**0.095**

Table 5. AUC, C-index, Brier score, and ECE computed at the 25th, 50th and 75th percentiles for the five models on the test set. The diagnosis codes are embedded using graph embedding. The best performing value for each evaluation metric is highlighted in bold.

Discussion

Use different models at different time points. We see the dependency of each model's performance on the time point at which it is evaluated. At a lower time point (e.g., at 25th percentile which is the 17-day time frame, and the 30-day time frame), DCM has the best discriminative power while SparseDeepWeiSurv is the best calibrated. For larger time frames (e.g., 50th and 75th percentiles), the performance of Cox PH improves to closely match SparseDeepWeiSurv for calibration, while the performance of DCM closely tracks the other two top performing methods, especially for the 50th time-quantile. This time dependency of the models' performance on the health care claims data may necessitate a different choice of model in quality of care issue detec-

30 day				
	AUC	C-index	Brier Score	ECE
Cox PH	0.827	0.821	0.032	0.056
C-mix	0.811	0.805	0.031	0.036
DeepSurv	0.811	0.805	0.033	0.056
DCM	**0.836**	**0.831**	0.032	0.054
SparseDeepWeiSurv	0.830	0.824	**0.029**	**0.009**

Table 6. AUC, C-index, Brier score, and ECE computed for 30-day readmission for the five models on the test set. The diagnosis codes are multi-hot encoded. The best performing value for each evaluation metric is highlighted in bold.

tion at different points in time. For example, suppose we are evaluating whether a readmission, which happens 100 days after its previous discharge, the choice of model used to evaluate the readmission should depend on which model has the best performance at $t = 100$ days.

In our statistical tests for significance, the differences in performance of the DCM and SparseDeepWeiSurv are non-significant at the 25th and 50th percentiles of the discriminative metrics. In addition, SparseDeepWeiSurv outperforms DCM and other methods in calibration metrics. In our task of identifying unusually early readmissions, calibration plays a more important role than discrimination. The downstream decision is based on the predicted likelihood of readmission at and before the date a patient of interest is being readmitted to determine if the readmission falls out of the acceptable likelihood threshold. We also note that the simpler Cox PH has strong performance in terms of discriminative power, comparable with the much more complex model SparseDeepWeiSurv. Therefore, Cox PH may be favored in tasks that focus only on the ranking of patients' readmission risk level.

DeepSurv, which removes the assumption of linearity in Cox PH and using a neural network to model the hazard function, consistently has worse performance than Cox PH across discrimination and calibration metrics. Further investigation is needed to understand why it underperforms CoX PH on our data.

There are some promising directions that we would like to explore as next steps. First, while our current approach is based solely on claims data, in future we would like to explore complementary data sources such as electronic health records through which we may be able to enrich our feature set to include lab results, radiology reports, etc. Second, since our models are built to allow for inference on patients with *any* medical condition (not restricted to one or a small set of medical conditions), we would like to investigate to what extent this relaxation compromises the models' performance in either discriminative power or calibration. It is also important to know which of the four metrics are the most reliable and relevant for this problem of detecting quality of care issues through hospital readmission prediction.

Limitations. Our study has several limitations. First, while we view quality of care issues through the lens of hospital readmissions (as do several prior studies ("Hospital readmission rate as an indicator of quality of care" section)), there are studies which have not found a strong link between readmissions and quality of care. For example, in Ref.[14] the authors show that less than one-fifth of urgent readmissions were potentially avoidable based physician reviews of patient files in a prospective study. In contrast to most prior studies (including[12]), we focus on survival modeling to predict the likelihood of all-cause readmissions that may not be urgent (i.e. within 30 days) based on claims data. Second, the claims data we use in our experiments do not have an indicator of mortality, a confounding factor for survival analysis. The right-censoring of our data accounts for both mortality and no readmissions within 3 years following the initial admission. Third, our analysis is based on patients enrolled in the Medicare program (who are aged 65 years or over, younger people with disabilities, and people with End Stage Renal Disease) and our findings may not apply to readmissions data from other demographics.

Conclusion

In this study, we frame the problem of identifying early readmissions following a discharge as a survival analysis problem, where we estimate the distribution over time to readmission after a discharge conditioned on the discharge's covariates. We evaluate five models both on the discriminative power and the calibration. We observe that Cox PH and SparseDeepWeiSurv, the top performing models, have comparable discrimination ability; but SparseDeepWeiSurv, which models the time to readmission as a Weibull distribution, is the better calibrated. DeepSurv, which removes the linearity assumption of Cox PH and replaces it with a more complex relationship modeled by a neural network, has worse performance than Cox PH. DCM, an extension of DeepSurv, also generally performs worse than DeepSurv on our dataset. We also find that representing the diagnosis codes with advanced embedding methods such as those from Node2Vec and BioBERT does not improve and, in some cases, worsens model performance.

Data availability

The raw datasets analysed during the current study are not publicly available in full due to licensing and contractual restrictions, but synthetic sample datasets are available from the corresponding author on reasonable request. The source dataset was redacted and anonymized by a team who are specialized in this process, following

the Safe Harbor method, Section 164.514(b)(2) of the HIPAA Privacy Rule. Deloitte holds contracts with various Medicare and Medicaid agencies through which it has access to this data. We cannot advice on conditions under which other researchers can access similar datasets. The CMS provides beneficiary-level health information to researchers which can be requested through https://www.cms.gov/research-statistics-data-and-systems/files-for-order/limiteddatasets. Implementations of the models we experimented with are the following: Cox PH https://lifelines.readthedocs.io/en/latest/index.html, C-mix https://github.com/SimonBussy/C-mix, DeepSurv https://github.com/jaredleekatzman/DeepSurv, DeepWeiSurv https://github.com/survml, DeepCoxMixture https://autonlab.org/auton-survival/.

Received: 18 October 2022; Accepted: 22 June 2023
Published online: 28 June 2023

References

1. Jencks, S. F., Williams, M. V. & Coleman, E. A. Rehospitalizations among patients in the Medicare fee-for-service program. *N. Engl. J. Med.* **360**, 1418–1428 (2009).
2. Lanièce, I. *et al.* Incidence and main factors associated with early unplanned hospital readmission among French medical inpatients aged 75 and over admitted through emergency units. *Age Ageing* **37**, 416–422 (2008).
3. Krumholz, H. M. *et al.* Readmission after hospitalization for congestive heart failure among Medicare beneficiaries. *Arch. Intern. Med.* **157**, 99–104 (1997).
4. Sharma, V. *et al.* Predicting 30-day readmissions in patients with heart failure using administrative data: A machine learning approach. *J. Card. Fail.* **28**, 710 (2021).
5. Hines, A. L. *et al. Conditions with the Largest Number of Adult Hospital Readmissions by Payer, 2011: Statistical brief# 172* (2014).
6. Canadian Institute for Health Information: All-Cause Readmission to Acute Care and Return to the Emergency Department. https://publications.gc.ca/collections/collection_2013/icis-cihi/H118-93-2012-eng.pdf.
7. Baillie, C. A. *et al.* The readmission risk flag: Using the electronic health record to automatically identify patients at risk for 30-day readmission. *J. Hosp. Med.* **8**, 689–695 (2013).
8. All-cause readmission to acute care and return to the emergency department. In *Health system performance.* (Canadian Institute for Health Information, 2012)
9. Boutwell, A. & Hwu, S. *Effective Interventions to Reduce Rehospitalizations: A Survey of the Published Evidence* (Institute for Healthcare Improvement, 2009).
10. Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Policy Changes and Fiscal Year 2022 Rates; Quality Programs and Medicare Promoting Interoperability Program Requirements for Eligible Hospitals and Critical Access Hospitals; Changes to Medicaid Provider Enrollment; and Changes to the Medicare Shared Savings Program. https://www.federalregister.gov/documents/2021/08/13/2021-16519/medicare-program-hospital-inpatient-prospective-payment-systems-for-acute-care-hospitals-and-the.
11. Wadhera, R. K. *et al.* Hospital revisits within 30 days after discharge for medical conditions targeted by the Hospital Readmissions Reduction Program in the United States: National retrospective analysis. *BMJ* **366**, 4563 (2019).
12. Kristensen, S. R., Bech, M. & Quentin, W. A roadmap for comparing readmission policies with application to Denmark, England, Germany and the United States. *Health Policy* **119**, 264–273 (2015).
13. Kansagara, D. *et al.* Risk prediction models for hospital readmission: A systematic review. *JAMA* **306**, 1688–1698 (2011).
14. Van Walraven, C. *et al.* Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ* **182**, 551–557 (2010).
15. Donzé, J. *et al.* Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. *JAMA Intern. Med.* **173**, 632–638 (2013).
16. Artetxe, A., Beristain, A. & Grana, M. Predictive models for hospital readmission risk: A systematic review of methods. *Comput. Methods Progr. Biomed.* **164**, 49–64 (2018).
17. Sushmita, S. *et al.* Predicting 30-day risk and cost of "all-cause" hospital readmissions. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence* (2016).
18. Mortazavi, B. J. *et al.* Analysis of machine learning techniques for heart failure readmissions. *Circ. Cardiovasc. Qual. Outcomes* **9**, 629–640 (2016).
19. Zhang, J. *et al. Hospital Readmission Prediction Using Swarm Intelligence-Based Support Vector Machines* 1522 (Institute of Industrial and Systems Engineers, 2013).
20. Min, X., Yu, B. & Wang, F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on COPD. *Sci. Rep.* **9**, 1–10 (2019).
21. Bussy, S. *et al.* Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework. *BMC Med. Res. Methodol.* **19**, 1–9 (2019).
22. Hempstalk, K. & Mordaunt, D. Improving 30-day readmission risk predictions using machine learning. In *Health Informatics New Zealand (HiNZ) Conference*, Vol. 2016 (2016).
23. Liu, W. *et al.* Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PLoS ONE* **15**, e0221606 (2020).
24. Frizzell, J. D. *et al.* Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches. *JAMA Cardiol.* **2**, 204–209 (2017).
25. Wang, L. *et al.* Predicting risk of hospitalization or death among patients with heart failure in the veterans health administration. *Am. J. Cardiol.* **110**, 1342–1349 (2012).
26. Betihavas, V. *et al.* An absolute risk prediction model to determine unplanned cardiovascular readmissions for adults with chronic heart failure. *Heart Lung Circ.* **24**, 1068–1073 (2015).
27. Padhukasahasram, B. *et al.* Joint impact of clinical and behavioral variables on the risk of unplanned readmission and death after a heart failure hospitalization. *PLoS ONE* **10**, e0129553 (2015).
28. Hao, S. *et al.* Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine healthcare information exchange. *PLoS ONE* **10**, e0140271 (2015).
29. Katzman, J. L. *et al.* DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 1–12 (2018).
30. Bennis, A., Mouysset, S. & Serrurier, M. Estimation of conditional mixture Weibull distribution with right censored data using neural network for time-to-event analysis. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I* 687–698 (Springer, 2020).
31. Bennis, A., Mouysset, S. & Serrurier, M. *DPWTE: A Deep Learning Approach to Time-to-Event Analysis using a Sparse Weibull Mixture Layer* (2021).

32. Lee, C., Zame, W., Yoon, J. & Van Der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In *Proc. AAAI Conference on Artificial Intelligence*, Vol. 32 (2018).
33. Nagpal, C., Li, X. & Dubrawski, A. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE J. Biomed. Health Inform.* **25**(8), 3163–3175 (2021).
34. Nagpal, C. *et al. Deep Cox Mixtures for Survival Regression* 674–708 (PMLR, 2021).
35. Ranganath, R., Perotte, A., Elhadad, N. & Blei, D. Deep survival analysis. In *Machine Learning for Healthcare Conference* 101–114 (PMLR, 2016).
36. Sukul, D. *et al.* Patterns of readmissions for three common conditions among younger US adults. *Am. J. Med.* **130**, 1220 (2017).
37. Krumholz, H. M. *et al.* An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. *Circ. Cardiovasc. Qual. Outcomes* **4**, 243–252 (2011).
38. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1721–1730 (2015).
39. International Classification of Diseases. *(ICD-10-CM/PCS) Transition—Background*. https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm.
40. CPT® (Current Procedural Terminology). https://www.ama-assn.org/amaone/cpt-current-procedural-terminology.
41. NPPES NPI Registry. https://npiregistry.cms.hhs.gov/.
42. CMS Specialty Codes/Healthcare Provider Taxonomy Crosswalk. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/MedicareProviderSupEnroll/downloads/taxonomy.pdf.
43. Zhang, Y. *et al.* BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* **6**, 52. https://doi.org/10.1038/s41597-019-0055-0 (2019).
44. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864 (2016).
45. Gorra, A. S. Using hospital outcomes to predict 30-day mortality among injured patients insured by medicare. *Arch. Surg.* **146**, 195. https://doi.org/10.1001/archsurg.2010.318 (2011).
46. Shebeshi, D. S., Dolja-Gore, X. & Byles, J. Unplanned readmission within 28 days of hospital discharge in a longitudinal population-based cohort of older Australian women. *Int. J. Environ. Res. Public Health* **17**, 3136. https://doi.org/10.3390/ijerph17093136 (2020).
47. Krumholz, H. M. *et al.* Do non-clinical factors improve prediction of readmission risk? Results from the Tele-HF study. *JACC Heart Fail.* **4**, 12–20 (2016).

Author contributions

N.Q.T. and G.G. prepared the dataset, extracted the features, and built and evaluated the models. D.L., J.H., N.P. and S.B. conceptualized, designed, and supervised the study. N.Q.T. prepared the first draft of the manuscript and S.B. prepared the revised manuscript. M.S. conducted additional experiments to address reviewers' comments. D.O. provided advise as a subject matter specialist in interpreting the data and results. E.B. helped with conducting the study and commented on the manuscript. All authors read and approved the manuscript.

Funding

The funding was provided by Deloitte & Touche LLP.

Competing interests

All authors are employees of Deloitte & Touche LLP.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-37477-3.

Correspondence and requests for materials should be addressed to S.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.