Improving the Calibration of Long-Term Predictions of Heart Failure Rehospitalizations using Medical Concept Embedding

Sunil Vasu Kalmady[1, 4], **Weijie Sun**[1, 2], **Justin Ezekowitz**[1], **Nowell Fine**[3], **Jonathan Howlett**[3], **Anamaria Savu**[1], **Russ Greiner**[2, 4], **Padma Kaul**[1]

[1] *Canadian VIGOUR Centre, Department of Medicine, University of Alberta, Alberta, Canada*

[2] *Department of Computing Science, University of Alberta, Alberta, Canada*

[3] *Libin Cardiovascular Institute of Alberta, University of Calgary, Alberta, Canada*

[4] *Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada*

KALMADY@UALBERTA.CA

Abstract

'Medical concept embedding' aims to provide vector representations of International Statistical Classification of Diseases (ICD) codes such that the relationship between two vectors mirrors the conceptual relationship between the two diagnoses or clinical interventions. Despite the growing interest in vector representations of clinical information in electronic health records (EHR), the utility of embedding methods has not been examined in the context of predicting individualized survival distributions (ISD). In this study, we apply ISD methods, specifically Cox-Proportional Hazards with Kalbfleisch-Prentice extension (CoxPH-KP) and Multi-task Logistic Regression (MTLR), to the task of predicting probability of Heart Failure (HF) rehospitalization or mortality, in a population-level database of 40,568 HF hospitalizations over the span of 8 years. Further, we compare performance of these ISD models with versus without code embeddings, that were learned in a temporally disjoint dataset of 229,359 all-cause hospitalizations. All our models show good discrimination in the validation dataset of 8,114 HF hospitalizations, with time-based concordance greater than 70% for every monthly intervals upto 8 years. Finally, we demonstrate that medical concept embedding does not always lead to improved model discrimination, but does improve model calibration, particularly over the longer time scales.

Keywords: Machine Learning, EHR, Embedding, Survival Analysis, Heart Failure

1. Introduction

Heart failure (HF) is a severe form of heart disease where the heart cannot pump enough blood to keep up with the body's needs. HF patients are at high risk of severe episodes and death. Moreover, patients who are hospitalized for HF are more likely to need to be hospitalized again, which poses high economic costs (Tran et al., 2016). Currently, there are no reliable models that identify which of these hospitalized HF patients are at high risk of unplanned rehospitalization or death.

It is challenging to build models that predict readmissions over long time scales using administrative databases because such clinical characterizations are often shallow or incomplete. Survival analyses, in particular, individual survival time distribution (ISD) methods, are well suited for this task as they can account for patient-level clinical heterogeneity and provide meaningful probabilistic estimates at several time points (Haider et al., 2018). ISD

also allows us to compute a specific patient's expected survival time. For example, ISD tools such as Cox-Proportional Hazards with Kalbfleisch-Prentice Extension (CoxPH-KP) (Kalbfleisch and Prentice, 2002) and Multi-task Logistic Regression (MTLR) (Yu et al., 2011) can be used to learn patient-specific survival curves based on patient attributes such demographic profile, comorbidities and medical interventions – see Figure 2. Here, for each time t, $S(t|x)$ is the probability that patient x will survive (not have an "event") at least until time t (where "event" means readmission to the hospital or death).

Administrative databases, or electronic health records (EHR) in general, systematically record clinical entities such as diagnoses and medical interventions of patients using international standard codes (WHO, 2016). Further, these datasets contain a sequence of hospital visits for each patient over time, where each visit can include multiple medical codes. We can encode the discrete codes from a finite set of choices as a multi-hot vector for each visit. However, these codes are often correlated – eg, ICD-10 code I50 (Heart Failure) and I48 (Atrial Fibrillation and Flutter) (Denaxas et al., 2018). The raw bit representation of multi-hot vector lacks the conceptual relationship between the codes, both semantically and in terms of the geometric distance between vectors. We may be able to address this problem with the use of embeddings, which translate large sparse vectors into a lower-dimensional space that implicitly embody semantic relationships. Since its first application in 2016, medical concept embeddings have been useful for predicting future medical codes (Choi et al., 2016b) or in analysis of patient similarity (Zhu et al., 2016). EHR-based embedding representations have been particularly successful in HF, and have led to promising results in predicting the risk of developing the disease in binary classification framework (Che et al., 2017; Denaxas et al., 2018; Choi et al., 2018). However, to best of our knowledge, the utility of using embeddings has not been examined in the context of survival (time to event) analysis or prediction of individualized survival curves in HF or any other medical condition.

Typically, clinical utility of predictive models that output probabilities are evaluated based on discrimination and calibration (Harrell Jr, 2015). Discrimination metrics, such as concordance, measures how well a model separates individuals into two classes, such as readmitted within 30 days versus not readmitted. On the other hand, calibration measures whether predicted probabilities agree with observed proportions. Calibration is particularly important in prognostic settings such as prediction of future readmission probability (Cook, 2008), since a poorly calibrated model might over- or under-estimate the readmission probability, even if it can still accurately classify individuals into classes –i.e., exhibit good discrimination (Figure 5 provides an illustrative example).

In this study, we developed and evaluated ISD models to predict HF rehospitalization or all cause mortality, with 40,568 HF hospitalizations in Canadian province of Alberta, over a 8 year period (2008 – 2016) using CoxPH-KP and MTLR methods. Further, we compared the model performance with and without the use of medical code embeddings, in terms of discrimination and calibration at monthly time points.

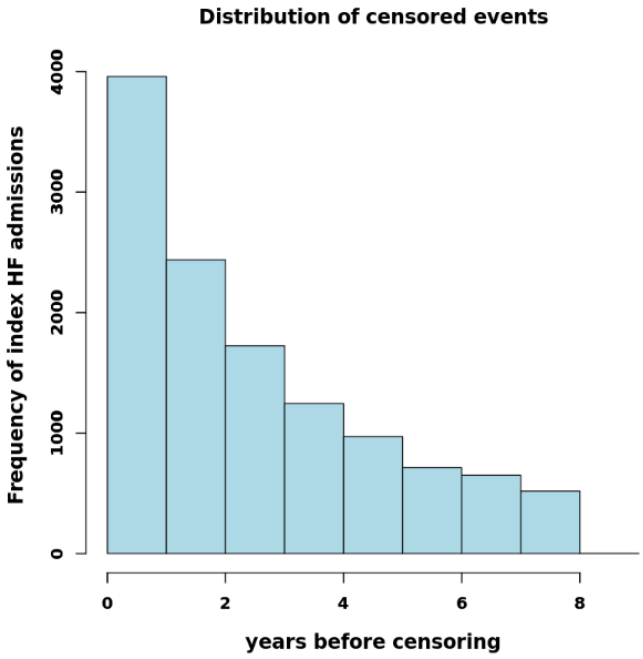

Figure 1: Frequency distribution of censored events

2. Methods

2.1. Data Sources

The province of Alberta has a single-payer, government funded health care system that provides universal access to over 4.3 million people for hospital, emergency department (ED), and physician services. We used de-identified data from administrative databases maintained by Alberta Health for the period of 2002 to 2016, including (1) the Discharge Abstract Database, which records for each hospital visit, the admission date, discharge date, most responsible diagnosis, and up to 25 other diagnoses, most responsible intervention, and up to 20 other interventions, special care units and physician specialities for all acute care hospitalizations; (2) the Ambulatory Care Database, which records all patient visits to hospital-based physicians' offices or EDs; (3) the Practitioner Claims Database, which tracks all physician claims for outpatient services; and (4) the Alberta Health Care Insurance Plan Registry, which tracks vital status of all residents. Lastly, demographic information such as sex, socio-economic status, urbanicity and ethnicity were collected from the Population registry. This study received ethics approval from the Health Ethics Research Board at the University of Alberta.

2.2. Feature Encoding

We used two types of encoding to represent the diagnosis / intervention ICD-10-CM (International Statistical Classification of Diseases and Related Health Problems - 10 - clinical modification) codes in each HF hospitalization, namely multi-hot and embedding. We partitioned the dataset based on time periods: Apr 2002 to Mar 2008, used to learn embedded

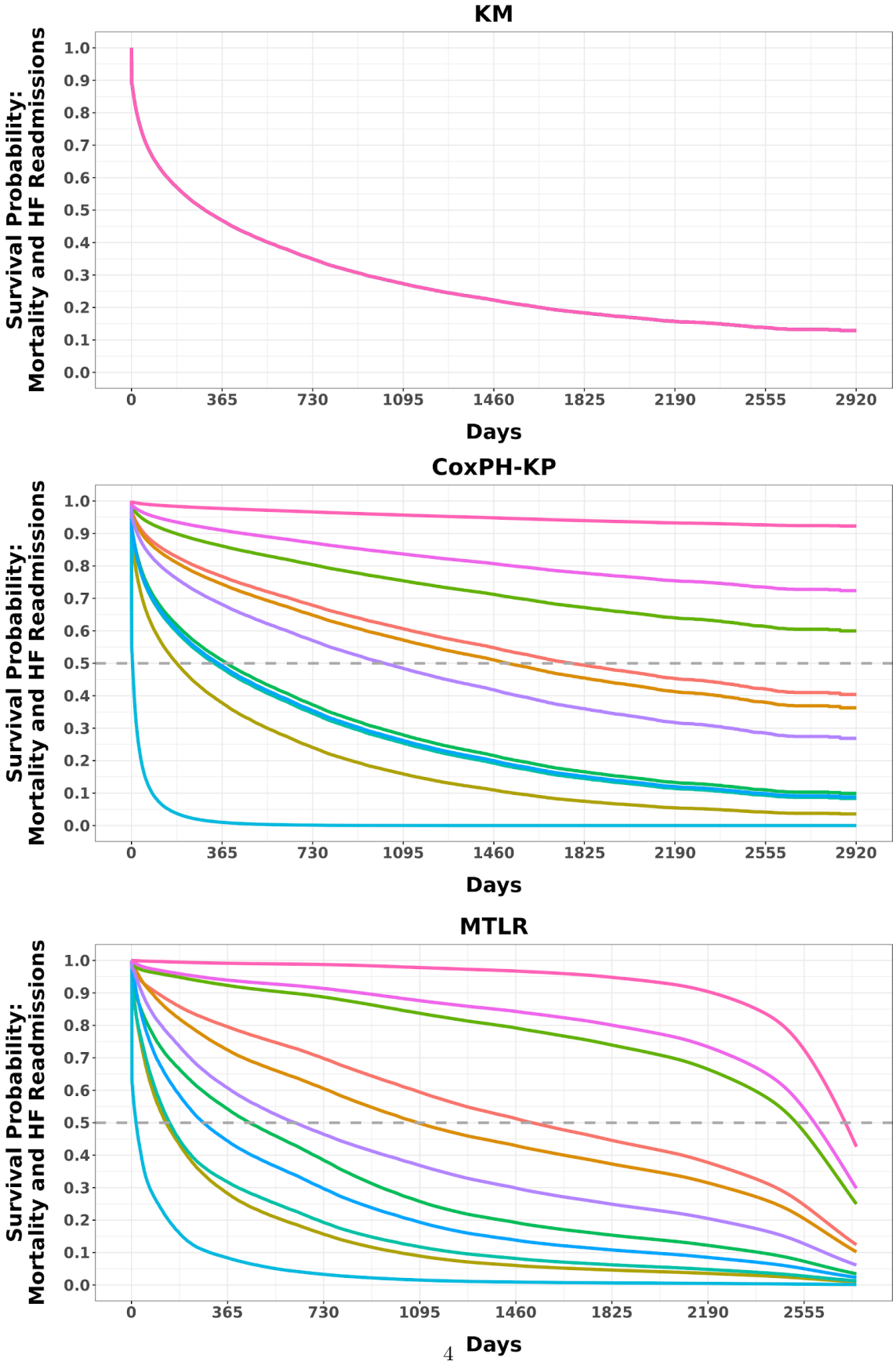

Figure 2: Representative survival curves for KM, CoxPH-KP and MTLR models

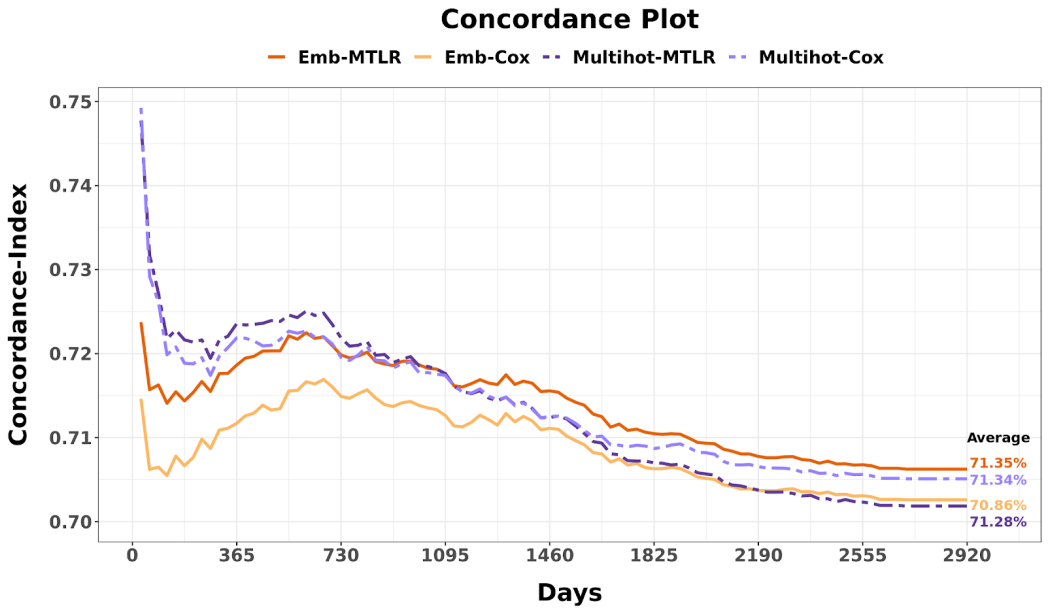

Figure 3: Concordance index measured at multiple time points and its average for the 4 evaluated models

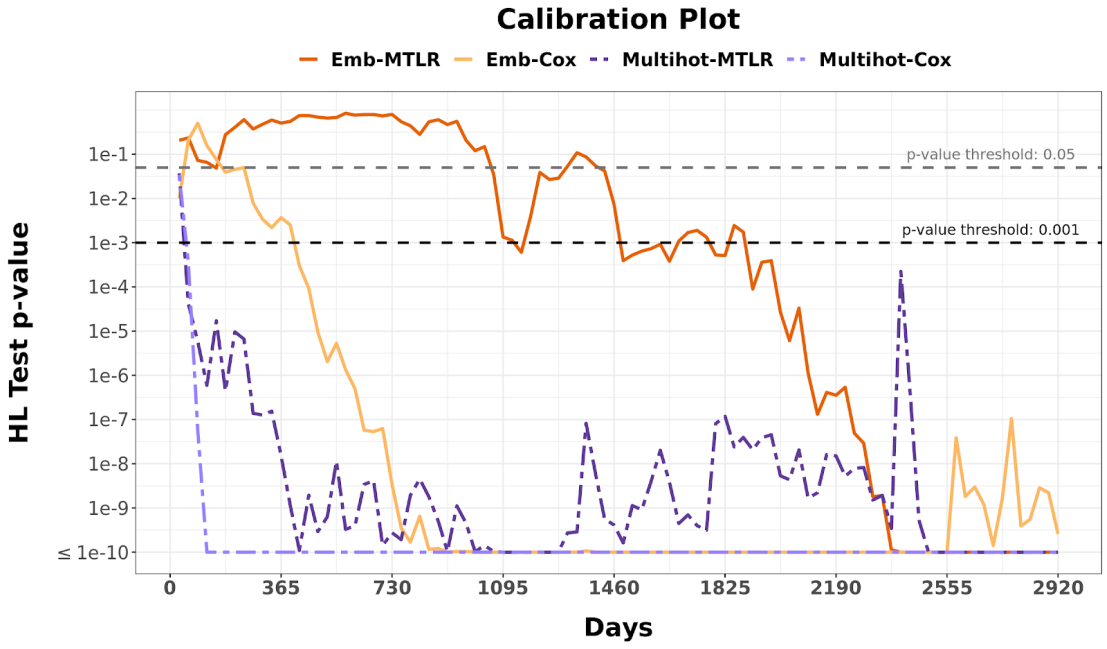

Figure 4: Calibration measured using HL test at multiple time points for the 4 evaluated models. Note that we consider the model to be calibrated if the p-value is greater than the threshold p-value – here, we consider 0.05 and 0.001.

representations and Apr 2008 to Mar 2016, used to learn and evaluate ISD models. In general, an ICD-10-CM code is 3 to 7 characters; we collapsed the less-frequent ICD codes (the ones ranked after the first 500) to their first 3 characters, which denotes the general category of disease. This procedure reduced the total of 12827 unique ICD codes to 3015 codes, which was then used for both kinds of encoding.

To derive the embeddings, we used inpatient records pertaining to all cause hospitalizations (total: 229,359) from 37,161 patients during a 6 year period (2002 - 2008) who had at least 2 visits. We then applied the Med2Vec architecture (Choi et al., 2016a), which learns representations at two levels: code-level and visit-level. Here, we trained neural networks to predict codes in neighboring visits (previous and next), given the codes in current visit (input and outputs are multi-hot 3015 length vectors). Med2Vec combines this objective function of minimizing the cross-entropy with another that tries to find representations for the codes that predict others in the same visit, by maximizing the log-likelihood of co-occurring code representations (see Appendix A for more details). We set the embedding size to 200.

2.3. Predictive Modeling

The goal of our model was to predict the time of next hospitalization with primary diagnosis of HF (ICD-10 Code: starting with 'I50') or death during the 8 year period (2008 - 2016), given an input of features pertaining to a particular (index) HF hospitalization of an individual patient. Our dataset contained 40,568 primary HF hospitalizations of 29,525 patients, where outcome was right censored for 30.1% of hospitalizations – that is, 30.1% were alive and never re-hospitalized during this interval (Figure 1). In addition to medical codes (3015 in multi-hot encoding; 200 in embedded encoding), we used the following attributes of index hospitalization as features : age, sex, socio-economic status, urbanicity, ethnicity, physician's specialty, types of critical care unit, occurrence of hospital transfer, duration of hospital stay, duration of special care stay, total duration of hospital stay in preceding year, time since last all-cause hospitalization, and the counts of clinical encounters prior to index hospitalization in the preceding year of following types: inpatient, outpatient, ED, physician office visits each with primary HF, secondary HF or any diagnosis. We performed feature selection based on significance (p-value < 0.1) of the variables in univariate CoxPH regression on the training dataset. We used CoxPH-KP and MTLR to learn ISD models over these selected features (see Appendix B, C for details) with and without embedding, namely: Emb-Cox, Emb-MTLR, multihot-Cox and multihot-MTLR.

2.4. Evaluation

We trained each model (Emb-Cox, Emb-MTLR, multihot-Cox and multihot-MTLR) using 4/5th of the dataset (n = 32,454), and then validated the learned model on the held out 1/5th (n = 8114) of dataset. This split was stratified such that the fraction of censored events and range of event times was roughly equal in the training and held out test set.

We evaluated the performance of our models using concordance and calibration, calculated at every time point corresponding to monthly (30 days) intervals up to 8 years (total: 96 time points). We computed concordance to measure discriminative ability of the model to categorize the hospitalizations as with / without outcome, while only considering pairs

of hospitalizations in different predefined bins [1] (Heagerty and Zheng, 2005). Here the risk score was given by the negative of the median survival time [2]. Similarly, we used calibration to measure the ability of the model to correctly estimate the probability of outcome at multiple time points, with D'Agostino-Nam translation to account for censored events, using Hosmer-Lemeshow (HL) goodness-of-fit test (Hosmer and Lemesbow, 1980). We also computed a summarised performance measure, namely the integrated brier score (Brier and Allen, 1951).

3. Results

Figure 2 shows survival curves for Kaplan-Meier (KM) model as well as Emb-Cox, Emb-MTLR models for 10 representative test hospitalizations. Since the KM model does not account for individual features, the survival distribution represents the population's overall survival curve. Next, note that all the CoxPH-KP curves have effectively the same shape, and do not cross, due to the proportional hazards assumption, whereas the curves for MTLR can cross.

Figures 3 and 4 show the concordance-index and calibration HL p-value for the 4 prediction models, over 96 monthly time points. With respect to concordance, we observed that all the 4 models range from 70 to 75% over all the time points (Figure 3). Embedding models were slightly more stable, and in particular Emb-MTLR shows lowest variance and highest average (although all 4 averages were very close to each other). Note that multi-hot models performed particularly well in early time points (time points $<$ 1 year).

For inferring the model calibration, in addition to conventional p-value threshold of 0.05, we also compared the models based on relaxed threshold of 0.001, given the increased power of HL test at large sample sizes (see horizontal dashed lines in Figure 4). We observed the Emb-MTLR was most calibrated among the evaluated models. It was well calibrated upto 3 years at 0.05 threshold, and upto 4 to 5 years at 0.001 threshold (Figure 4). The second most calibrated model was Emb-Cox which showed good calibration upto 6 months to a year. On the other hand, both multi-hot models (multihot-Cox and multihot-MTLR) were not well calibrated except for first time point (30-day prediction), but only at 0.001 threshold.

Lastly, the integrated Brier score was 14.97, 14.82, 15.01 and 14.87 for Emb-Cox, Emb-MTLR, multihot-Cox and multihot-MTLR models respectively. Again, Emb-MTLR showed the best score [3]. However, it should be noted that depending on the clinical context, a more accurate survival model at shorter duration, such as at 30 days, might be more useful than a more stable model at longer duration. Additionally, given the long time span of the data, any temporal changes in adoption and definitions in ICD-10 codes could potentially effect the embeddings and prediction performance. In summary, as expected, we observed a drop in performance with both discrimination and calibration over longer range predictions. All 4 models evaluated were comparable in terms of overall concordance, however models trained with medical code embeddings showed more stable predictions and better calibration over the longer time scales compared to models trained with multi-hot encoded features.

1. e.g. at 30 day cut point, Bin_1=[0, 30] and Bin_2=[31, 2920]
2. Median survival time is where the survival curve crosses 50% probability, see horizontal line in Figure 2
3. For Brier score, lower score indicates better performance. Baseline score is 0.25.

References

Glenn W Brier and Roger A Allen. Verification of weather forecasts. In *Compendium of meteorology*, pages 841–848. Springer, 1951.

Zhengping Che, Yu Cheng, Zhaonan Sun, and Yan Liu. Exploiting convolutional neural network for risk prediction with medical feature embedding. *arXiv preprint arXiv:1701.07474*, 2017.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504, 2016a.

Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*, 2016b.

Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in neural information processing systems*, pages 4547–4557, 2018.

Nancy R Cook. Statistical evaluation of prognostic versus diagnostic models: beyond the roc curve. *Clinical chemistry*, 54(1):17–23, 2008.

Spiros Denaxas, Pontus Stenetorp, Sebastian Riedel, Maria Pikoula, Richard Dobson, and Harry Hemingway. Application of clinical concept embeddings for heart failure prediction in uk ehr data. *arXiv preprint arXiv:1811.11005*, 2018.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *arXiv preprint arXiv:1811.11347*, 2018.

Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

David W. Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):1043–1069, 1980.

JD Kalbfleisch and RL Prentice. The statistical analysis of failure time data 2002 new york. *NY Wiley Crossref*, 2002.

Yi ping Weng. Baseline survival function estimators under proportional hazards assumption. 2007.

Dat T Tran, Arto Ohinmaa, Nguyen X Thanh, Jonathan G Howlett, Justin A Ezekowitz, Finlay A McAlister, and Padma Kaul. The current and future financial burden of hospital admissions for heart failure in canada: a cost analysis. *CMAJ open*, 4(3):E365, 2016.

WHO. *International statistical classification of diseases and related health problems*, volume 2. World Health Organization, 2016.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, pages 1845–1853, 2011.

Zihao Zhu, Changchang Yin, Buyue Qian, Yu Cheng, Jishang Wei, and Fei Wang. Measuring patient similarities via a deep architecture with medical concept embedding. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 749–758. IEEE, 2016.

Appendix

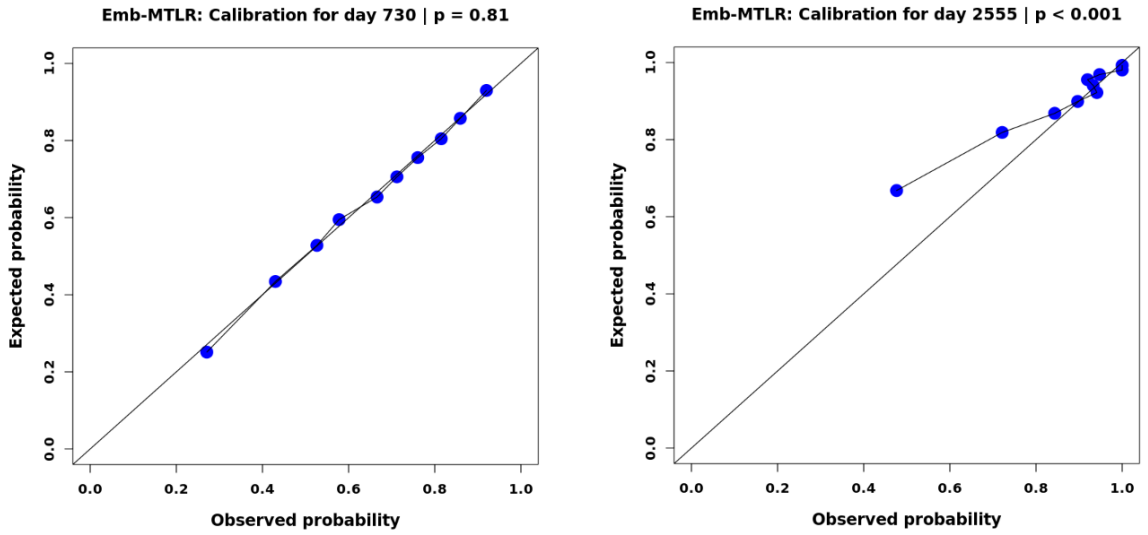

Figure 5: The bins of observed and expected probabilities associated with two calibration computations of the Emb-MTLR model applied for 2 year [left] and 7 year [right] time points. Note that left is well calibrated, while right is not calibrated, however both show good discrimination (concordance index $> 70\%$)

Appendix A. Overview of Med2Vec

Choi, Edward and Bahadori, Mohammad Taha et al. provides the Med2Vec algorithm (Choi et al., 2016a). Here we provide an brief summary. To quote the authors:

We can denote the set of all medical codes $c_1, c_2, ..., c_{|\mathcal{C}|}$ in our EHR dataset by $\mathcal{C}$ with size $|\mathcal{C}|$. EHR data for each patient is in the form of a sequence of visits $V_1, ..., V_T$ where each visit contains a subset of medical codes $V_T \in \mathcal{C}$. The goal of Med2Vec is to learn two types of representations:

Code representations: First goal is to learn an embedding function $f_C : \mathcal{C} \mapsto \mathbf{R}_+^m$ that maps every code in the set of all medical codes $\mathcal{C}$ to non-negative real-valued vectors of dimension m.

Visit representations: Second goal is to learn another embedding function $f_\mathcal{V} : \mathcal{V} \mapsto \mathbf{R}^n$ that maps every visit (a set of medical codes) to a real-valued vector of dimension n. The set $\mathcal{V}$ is the power set of the set of codes $\mathcal{C}$.

Given a visit V_t, a multi-layer perceptron generates the corresponding visit representation v_t. First, visit $\mathcal{V}_t$ is represented by a multi-hot vector $x_t \in \{0, 1\}^{|\mathcal{C}|}$ where the ith entry is 1 only if $c_i \in V_t$. Then x_t is converted to an intermediate visit representation $u_t \in \mathbf{R}^m$

Appendix

as follows

$$\boldsymbol{u}_t = ReLU(\mathbf{W}_c \boldsymbol{x}_t + \boldsymbol{b}_c) \tag{1}$$

using the code weight matrix $\mathbf{W}_c \in \mathbf{R}^{m \times |\mathcal{C}|}$ and the bias vector $\mathbf{b}_c \in \mathbf{R}^m$.

The final visit representation $\mathbf{v}_t \in \mathbf{R}^n$ is created as follows,

$$\boldsymbol{v}_t = ReLU(\mathbf{W}_v \boldsymbol{u}_t + \boldsymbol{b}_v)$$

using the visit weight matrix $\mathbf{W}_v \in \mathbf{R}^{n \times m}$ and the bias vector $\mathbf{b}_v \in \mathbf{R}^n$, where n is the predefined size of the visit representation. (Note that unlike the original Med2Vec, we did not use demographic information for learning the embeddings)

Given a visit representation $\mathbf{v}_t$, a softmax classifier predicts the medical codes of the visits within a context window (previous and next events) by minimizing the cross entropy error as follows:

$$\begin{aligned} \min_{\boldsymbol{W}_s, \boldsymbol{b}_s} \frac{1}{T} \sum_{t=1}^{T} \sum_{-w \le i \le w, i \neq 0} & -(\boldsymbol{x}_{t+i})^\top \log \hat{y}_t \\ & -(1 - \boldsymbol{x}_{t+i})^\top \log (1 - \hat{y}_t), \end{aligned} \tag{2}$$

where

$$\hat{y}_t = \frac{\exp (\boldsymbol{W}_s \boldsymbol{v}_t + \boldsymbol{b}_s)}{\sum_{j=1}^{|\mathcal{C}|} \exp (\boldsymbol{W}_s[j,:] \boldsymbol{v} + \boldsymbol{b}_s[j])}$$

where $\mathbf{W}_s \in \mathbf{R}^{|\mathcal{C}| \times n}$ and $\boldsymbol{b}_s \in \mathbf{R}^{|\mathcal{C}|}$ are the weight matrix and bias vector for the softmax classifier, w the predefined context window size, exp the element-wise exponential function, and $\mathbf{1}$ denotes an all one vector.

The code representations to be learned is denoted as a matrix $\boldsymbol{W}'_c = ReLU(\boldsymbol{W}_c) \in \mathbf{R}^{m \times |\mathcal{C}|}$. From a sequence of visits $V_1, V_2, ..., V_T$, the code-level representations can be learned by maximizing the following log-likelihood,

$$\min_{\boldsymbol{W'}_c} \frac{1}{T} \sum_{t=1}^{T} \sum_{i: c_i \in V_t} \sum_{j: c_j \in V_t, j \neq i} \log p(c_j | c_i), \tag{3}$$

where

$$p(c_j | c_i) = \frac{\exp (\boldsymbol{W'}_c)[:, j]^\top \boldsymbol{W'}_c[:, i]}{\sum_{k=1}^{|\mathcal{C}|} \exp (\boldsymbol{W'}_c[:, k]^\top \boldsymbol{W'}_c[:, i])} \tag{4}$$

Appendix

Finally, the single unified framework can be obtained by adding the two objective functions 2 and 3 as follows,

$$\begin{aligned}\underset{\boldsymbol{W}_{c,v,s},\boldsymbol{b}_{c,v,s}}{\arg\min}\ \frac{1}{T}\sum_{t=1}^{T}\Bigg\{ &-\sum_{i:c_i\in V_t}\sum_{j:c_j\in V_t, j\neq i}\log p(c_j|c_i) \\ &+\sum_{-w\leq k\leq w,k\neq 0} -\boldsymbol{x}_{t+k}^{\top}\log\hat{y}_t-(\mathbf{1}-\boldsymbol{x}_{t+k})^{\top}\log(\mathbf{1}-\hat{y}_t)\Bigg\}\end{aligned} \tag{5}$$

By combining the two objective functions, Med2Vec learns both code representations and visit representations from the same source of patient visit records, exploiting both intra-visit co-occurrence information as well as inter-visit sequential information at the same time.

Appendix B. Overview of MTLR

Consider modeling the probability of survival of patients at each of a vector of time points $\tau = [t_1, t_2, \ldots, t_m]$. In our study τ is 96 monthly intervals from 1 month up to 8 years. We can set up a series of logistic regression models: For each patient, represented as $\boldsymbol{x}$,

$$P(T \geq t_i \,|\, \boldsymbol{x}) = (1+\exp(\boldsymbol{\theta_i}\cdot\boldsymbol{x}))^{-1} \tag{6}$$

where $\boldsymbol{\theta_i}$ are the time-specific feature vectors. While the input features $\boldsymbol{x}$ stay the same for all these classification tasks, the binary labels $y_i = [T \geq t_i]$ can change depending on the threshold t_i.% We encode the survival time d of a patient as a binary sequence $y = y(d) = (y_1, y_2, \ldots, y_m)$, where $y_i = y_i(d) \in \{0, 1\}$ denotes the survival status of the patient at time t_i, so that $y_i = 0$ (no death or readmission event yet) for all i with $t_i < d$, and $(y_i = 1)$ (death) for all i with $t_i \geq d$.

Here there are $m + 1$ possible legal sequences of the form $(0, 0, \ldots, 1, 1, \ldots, 1)$, including the sequence of all 0's and the sequence of all 1's. The probability of observing the survival status sequence $y = (y_1, y_2, \ldots, y_m)$ can be represented as:

$$P_\Theta(Y{=}(y_1, y_2, \ldots, y_m) \mid \boldsymbol{x}) = \frac{\exp(\sum_{i=1}^{m} y_i \times \theta_i \cdot \boldsymbol{x})}{\sum_{k=0}^{m}\exp(f_\Theta(\boldsymbol{x}, k))}, \tag{7}$$

where $\Theta = (\theta_1, \ldots, \theta_m)$, and $f_\Theta(\boldsymbol{x}, k) = \sum_{i=k+1}^{m}(\theta_i \cdot \boldsymbol{x})$ for $0 \leq k \leq m$ is the score of the sequence with the event occurring in the interval $[t_k, t_{k+1})$ before taking the logistic transform, with the boundary case $f_\Theta(\boldsymbol{x}, k) = 0$ being the score for the sequence of all '0's. Given a dataset of n patients $\{\boldsymbol{x_r}\}$ with associated time of events $\{d_r\}$, we find the optimal parameters (for the MTLR model) (Θ^*) as

$$\begin{aligned}\Theta^* \ = \ \arg\max_{\Theta}\sum_{r=1}^{n}\Big[\sum_{i=1}^{m} y_j(d_r)(\theta i\cdot\boldsymbol{x_r}) \\ -\log\sum_{k=0}^{m}\exp f_\Theta(\boldsymbol{x_r}, k)\Big] - \frac{C}{2}\sum_{j=1}^{m}\|\theta_j\|^2\end{aligned} \tag{8}$$

where the C (for the regularizer) is found by an internal cross-validation process. There are many details here – to insure that the survival function starts at 1.0, and decreases monotonically and smoothly until reaching 0.0 for the final time point; to deal appropriately with censored patients; to decide how many time points to consider (m); and to minimize the risk of overfitting (by regularizing), and by selecting the relevant features.

The paper by Yu et al. provides the details. C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In NIPS, 2011(Yu et al., 2011).

Appendix C. Overview of CoxPH-KP

The Cox proportional-hazards (Cox-PH) model is extremely common in the survival analysis literature. Cox-PH models the hazard function of patients, where hazard is interpreted as the risk of event at any given time, specifically for a time t and covariates $\boldsymbol{x}$ the hazard function is defined as

$$h(t \,|\, \boldsymbol{x}) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \;\mid\; T \geq t, \boldsymbol{x})}{\Delta t}. \tag{9}$$

Cox-PH models this hazard function in terms of a baseline hazard, $h_0(t)$, equal for all patients and scaled by a learned, individual risk depending on features, that is, $h(t|\boldsymbol{x}) = h_0(t)\, e^{\boldsymbol{w}^T \boldsymbol{x}}$. This is why the model is termed the proportional hazards model, the hazard for one patient is proportional to the hazard of all other patients, by a scale depending on individual features. To learn the values for $\boldsymbol{w}^T$ one does not need to know the baseline hazard function. Suppose a single patient died at time t_j. The probability that patient i was the one who died is given by

$$\begin{aligned} \frac{h(t_j \,|\, \boldsymbol{x_i})}{\sum_{k \in R(t_j)} h(t_j \,|\, \boldsymbol{x_k})} &= \frac{h_0(t_j)\, e^{\boldsymbol{w}^T \boldsymbol{x_i}}}{\sum_{k \in R(t_j)} h_0(t_j)\, e^{\boldsymbol{w}^T \boldsymbol{x_k}}}, \\ &= \frac{e^{\boldsymbol{w}^T \boldsymbol{x_i}}}{\sum_{k \in R(t_j)} e^{\boldsymbol{w}^T \boldsymbol{x_k}}}, \end{aligned} \tag{10}$$

where $R(t_j)$ is the set of patients still alive (at risk) at time t_j. Thus the likelihood for $\boldsymbol{w}^T$ is defined as

$$L\left(\boldsymbol{w}^T\right) = \prod_j \frac{e^{\boldsymbol{w}^T \boldsymbol{x_{i(j)}}}}{\sum_{k \in R(t_j)} e^{\boldsymbol{w}^T \boldsymbol{x_k}}}, \tag{11}$$

where subscript $i(j)$ is interpreted as it is patient i who had an event at time j. By this formulation of the likelihood equation, we see that feature weights are independent of time. While we are able to estimate feature weights without specifying the baseline hazard function, we cannot determine the survival distribution. A number of methods exist for calculating the the baseline hazard and therefore the survival distribution including the Kalbfleisch-Prentice estimator which we use in this work. The thesis "Baseline Survival Function Estimator under Proporitional Hazards Assumption" provides a nice discussion of the Breslow and Kalbfleisch-Prentice estimators (ping Weng, 2007).