Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Proc (IEEE Int Conf Healthc Inform). Author manuscript; available in PMC 2022 May 06.

Published in final edited form as:

Proc (IEEE Int Conf Healthc Inform). 2021 August ; 2021: 497–498. doi:10.1109/ichi52183.2021.00088.

An empirical study of using radiology reports and images to improve ICU-mortality prediction

Mingquan Lin[*,§], **Song Wang**[†,§], **Ying Ding**[†], **Lihui Zhao**[‡], **Fei Wang**[*], **Yifan Peng**[*]

[*]Department of Population Health Sciences, Weill Cornell Medicine, New York, USA

[†]Cockrell School of Engineering, The University of Texas at Austin, Austin, USA

[‡]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, USA

Abstract

The predictive Intensive Care Unit (ICU) scoring system plays an important role in ICU management for its capability of predicting important outcomes, especially mortality. There are many scoring systems that have been developed and used in the ICU. These scoring systems are primarily based on the structured clinical data contained in the electronic health record (EHR), which may suffer the loss of the important clinical information contained in the narratives and images. In this work, we build a deep learning based survival prediction model with multi-modality data to predict ICU-mortality. Four sets of features are investigated: (1) physiological measurements of Simplified Acute Physiology Score (SAPS) II, (2) common thorax diseases pre-defined by radiologists, (3) BERT-based text representations, and (4) chest X-ray image features. We use the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset to evaluate the proposed model. Our model achieves the average C-index of 0.7847 (95% confidence interval, 0.7625–0.8068), which substantially exceeds that of the baseline with SAPS-II features (0.7477 (0.7238–0.7716)). Ablation studies further demonstrate the contributions of pre-defined labels (2.12%), text features (2.68%), and image features (2.96%). Our model achieves a higher average C-index than the traditional machine learning methods under the same feature fusion setting, which suggests that the deep learning methods can outperform the traditional machine learning methods in ICU-mortality prediction. These results highlight the potential of deep learning models with multimodal information to enhance ICU-mortality prediction. We make our work publicly available at https://github.com/bionlplab/mimic-icu-mortality.

Index Terms—

Mortality prediction; Deep learning; Multimodal fusion

I. Introduction

Predictive ICU scoring systems are the measures of disease severity that are used to predict outcomes, typically mortality, of patients in the intensive care unit (ICU) [1]. The scoring

[§]Lin and Wang contribute equally in this paper.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

systems such as Simplified Acute Physiology Score (SAPS) II [2] are primarily based on the structured clinical data, which are frequently documented in electronic health record (EHR). In this paper, we first build the clinical prediction models that will predict ICU-mortality using the SAPS-II risk factors such as demographics, vital signs, and lab tests. These measurements were obtained in the first 24 hours of ICU admission. We then enrich the model with multimodal features extracted from radiology reports and chest X-rays. The radiology imaging and reading were studied in the first 24 hours. We hypothesize that including free texts and images provides better predictions of ICU-mortality than including clinical measurements alone. Experiments on the MIMIC-IV dataset [3] show that our multimodal models are substantially more accurate than the unimodal ones.

II. Method

In this study, we use one of the most popular survival analysis models, the Cox model [4], where the survival function is assumed to be

$$S_i(t|x_i) = S_0(t)^{e^{\psi(x_i)}}. \quad (1)$$

In this model, $S_0(t)$ is the baseline survival function that describes the risk for individuals with $x_i = 0$ and $\psi(x_i) = x_i\beta$ being the relative risk based on the covariants. Note that $S_0(t)$ is shared by all patients at time t. It is NOT associated with any individual covariants. The effect of the covariate values x_i on the survival function is to raise it to a power given by the relative risk. In this paper, we expand $\psi(x_i)$ by introducing a deep neural network with the fusion features from multiple sources: SAPS-II risk factors x_{saps}, text features x_{text}, and imaging features x_{img} (Figure 1). Our model is called DeepSurv-based model.

III. Experiments and Result

We use the MIMIC-IV dataset (Medical Information Mart for Intensive Care IV) to evaluate the proposed model [3]. MIMIC-IV is a de-identified clinical database composed of 382,278 patients admitted in the ICUs at Beth Israel Deaconess Medical Center. Of those, we excluded patients who had no CXR studies before the measurements have been completed and resulted in the SAPS-II score. Therefore, there are in total 9,928 patients included in this study. Out of these patients, 2,213 patients (22%) were deceased in the ICU. We use the C-index to assess the accuracy of our models. We use 200 bootstrap samples to obtain a distribution of the C-index and report the 95% confidence intervals. For each bootstrap experiment, we sample n patients with replacement from the whole set of n patients. We then split the sampled set into training (70%), validation (10%), and test (20%) sets.

We obtain the SAPS-II scores using the scripts in the MIMIC-IV repository[1]. The text embeddings are extracted using BlueBERT [5], which was pre-trained on the PubMed

[1] https://github.com/MIT-LCP/mimic-iv

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

abstracts and MIMIC-III notes. We use pycox[2], scikit-survival [6], and PyTorch to implement the framework.

We first compare the results of the baseline ICU scoring model and our models with six different feature settings as shown in Table I. The SAPS-II score is an integer point score between 0 and 163 directly obtained from the MIMIC-IV website. The SAPS-II risk factors model is trained using the 15 routine physiological measurements. SAPS-II risk factors + labels model and SAPS-II risk factors + transformer features are trained using 15 routine physiological measurements respectively combined with 14 thorax disease labels, transformer-based features. The SAPS-II risk factors + GCN features model is enriched with the GCN-based features. The SAPS-II risk factors + Image features model is enriched with chest X-ray image features. The multimodal features model is trained using SAPS-II risk factors with the combination of text features and chest X-ray image features using early average fusion.

We then compare the performances of the conventional machine learning model and deep learning model: CoxPH [6] and DeepSurv-based model. Table II shows the results for both models with two feature settings.

IV. Future work

There are three tasks we plan to do in the future. First, we plan to use joint fusion in the future to propagate the loss back to the feature extraction modules during training, which may improve the representation learning performance. Second, We will explore other domain knowledge and try different ways of incorporating knowledge graph into ICU-mortality prediction. Third, we plan to employ the longitudinal EHR to assist predicting ICU-mortality.

Acknowledgment

This project was supported by National Library of Medicine under award number 4R00LM013001 and Amazon Machine Learning Grant.

References

[1]. Lipshutz AK, Feiner JR, Grimes B, and Gropper MA, “Predicting mortality in the intensive care unit: a comparison of the university health consortium expected probability of mortality and the mortality prediction model iii,” Journal of intensive care, vol. 4, no. 1, pp. 1–8, 2016.

[2]. Le Gall J-R, Lemeshow S, and Saulnier F, “A new simplified acute physiology score (saps ii) based on a european/north american multicenter study,” JAMA, vol. 270, no. 24, pp. 2957–2963, 1993.

[3]. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, and Mark R, “MIMIC-IV.”

[4]. Cox DR, “Regression models and life-tables,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no. 2, pp. 187–202, 1972.

[5]. Peng Y, Yan S, and Lu Z, “Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets,” in Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019), 2019, pp. 58–65.

[2] https://github.com/havakv/pycox

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[6]. Pölsterl S, "scikit-survival: A library for time-to-event analysis built on top of scikit-learn," Journal of Machine Learning Research, vol. 21, no. 212, pp. 1–6, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-729.html

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

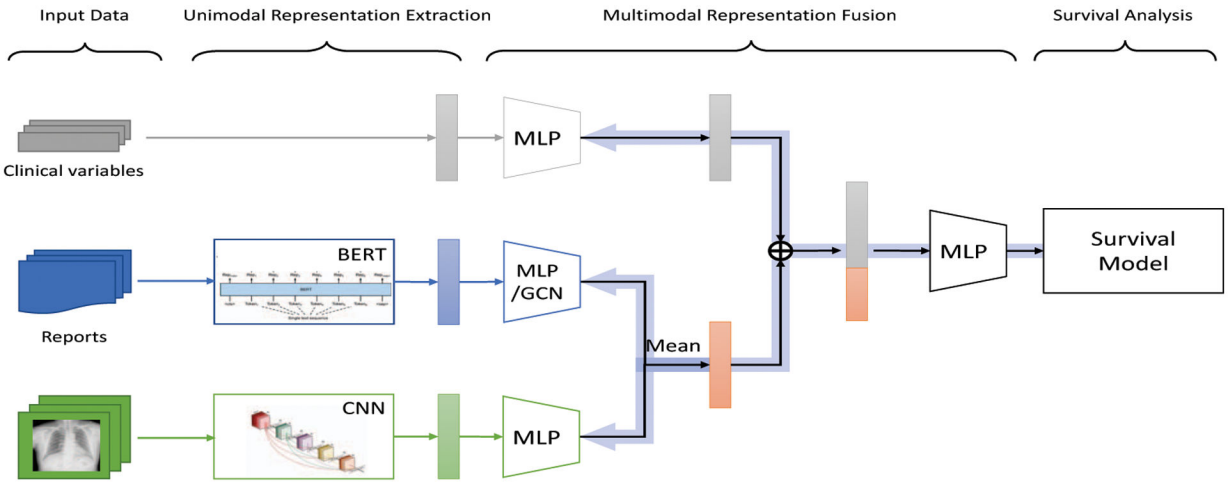

Fig. 1:
Multimodal feature fusion network.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table I:

C-index comparisons of the models using different sets of features.

Model	C-index (95% CI)
SAPS-II scores (ICU scoring baseline)	0.7477 (0.7238−0.7716)
SAPS-II risk factors	0.7555 (0.7220−0.7890)
SAPS-II risk factors + labels	0.7689 (0.7430−0.7948)
SAPS-II risk factors + transformer features	0.7733 (0.7498−0.7968)
SAPS-II risk factors + GCN features	0.7745 (0.7486−0.8004)
SAPS-II risk factors + Image features	0.7757 (0.7522−0.7992)
Multimodal features	**0.7847 (0.7625−0.8068)**

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table II:

The C-index results of the conventional machine learning models and the deep learning models trained and tested on the entire dataset.

Model		C-index (95% CI)
SAPS-II risk factors	CoxPH	0.7527 (0.7270−0.7784)
	DeepSurv-based	**0.7555 (0.7220−0.7890)**
SAPS-II risk factors	CoxPH	0.7643 (0.7392−0.7894)
+ labels	DeepSurv-based	**0.7689 (0.7430−0.7948)**