Leveraging Deep Representations of Radiology Reports in Survival Analysis for Predicting Heart Failure Patient Mortality

Hyun Gi Lee[1] **Evan Sholle**[2] **Ashley Beecy**[3] **Subhi Al'Aref**[4] **Yifan Peng**[1]
[1]Department of Population Health Sciences, Weill Cornell Medicine
[2]Information Technologies and Services, Weill Cornell Medicine
[3]Department of Medicine, Weill Cornell Medicine
[4]Department of Internal Medicine, UAMS College of Medicine

Abstract

Utilizing clinical texts in survival analysis is difficult because they are largely unstructured. Current automatic extraction models fail to capture textual information comprehensively since their labels are limited in scope. Furthermore, they typically require a large amount of data and high-quality expert annotations for training. In this work, we present a novel method of using BERT-based hidden layer representations of clinical texts as covariates for proportional hazards models to predict patient survival outcomes. We show that hidden layers yield notably more accurate predictions than predefined features, outperforming the previous baseline model by 5.7% on average across C-index and time-dependent AUC. We make our work publicly available at https://github.com/bionlplab/heart_failure_mortality.

1 Introduction

Survival analysis estimates the expected time until an event of interest occurs (Ranganath et al., 2016). In clinical research, it is used to understand the relationship between prognostic covariates (e.g., age and treatment) and patient survival time for important use cases such as predicting mortality of heart failure patients and providing management recommendations for intensive care units during a public health crisis like the COVID-19 pandemic (Pandey et al., 2020; Sprung et al., 2020; Nielsen et al., 2019).

Clinical texts such as radiology reports contain rich information about patients that is used to diagnose disease, plan treatments and monitor progress. It also contains the high-level reasoning of human experts that requires years of knowledge accumulation and professional training (Langlotz, 2015). Despite their clinical relevance, it is challenging to use them in survival analysis since they are largely unstructured. Automatic labelers are often unable to capture detailed information to distinguish between patients, especially ones with similar conditions, because they mostly rely on a small set of manually selected labels (Lao et al., 2017). Developing methods for accessing the critical information embedded in unstructured clinical texts holds the potential to meaningfully benefit clinical research.

To bridge this gap, we propose a deep learning method to predict the survival probability of heart failure (HF) patients based on the high-dimensional feature representations of their radiology reports. Concretely, we extract hidden features from the texts with BERT-based (Devlin et al., 2019) models and apply a recurrent neural network (RNN) to model sequences of reports and estimate the log-risk function for the overall mortality prediction. This approach can encapsulate more textual information than hand-crafted features and incorporate higher-order temporal information from report sequences. We find that our model improves on average 5.7% in both C-index and time-dependent AUC without requiring additional expert annotations.

We make three contributions through this work: (1) present a novel survival analysis model to leverage feature representations from clinical texts, (2) demonstrate that our model outperforms the ones dependent on predefined expert features and that this approach can generalize across various biomedical and clinical BERT models, and (3) make our work publicly available for reproduction by others.

2 Related Work

Due to the lack of expert annotations, earlier automatic labelers mostly use predefined linguistic patterns to extract relevant information. NegEx (Chapman et al., 2001) is a regular expression algorithm that identifies observations based on specified phrases. NegBio (Peng et al., 2018) uses universal dependencies and subgraph matching in addition to regular expressions. CheXpert (Irvin et al., 2019) extends NegBio by adding rules to extract, classify,

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4533–4538
June 6–11, 2021.

and aggregate mentions to improve performance. While they typically achieve a high precision, they suffer from a low recall because of their limited rules.

BERT (Devlin et al., 2019) is a transformer-based method that extracts feature representations of unlabeled text that are effective for transfer learning across various NLP tasks. It is adapted to a wide range of domains, including biomedical and clinical domains (Lee et al., 2020; Alsentzer et al., 2019; Peng et al., 2019). Recently, BERT models have been applied to labeling radiology reports. CheXbert (Smit et al., 2020) and CheXpert++ (McDermott et al., 2020) train on silver-standard datasets created with a rule-based labeler, CheXpert. Although they outperform rule-based labelers, these approaches need a curated training corpus which can be costly to obtain and error-prone. Furthermore, their labels are still limited and can miss critical information from the reports.

Regarding survival analysis, the Cox proportional hazards model (CPH) is widely adopted as it can deal with censored data and evaluate the prognostic values of covariates simultaneously (Cox, 1972). DeepSurv (Katzman et al., 2018) and DeepHit (Lee et al., 2018) are more contemporary methods that use deep neural networks to model more complex, nonlinear relationships of predictor variables. RNN-SURV (Giunchiglia et al., 2018) and DRSA (Ren et al., 2019) model time-variant, sequential patterns from predictor variables. To the best of our knowledge, the compatibility of these models and high-dimensional features as covariates has not been tested.

Automatic extraction tools enable survival analysis to incorporate textual information from clinical texts. Pandey et al. (2020) used a convolutional neural network to extract findings from radiology reports of heart failure patients and predict all-cause mortality with CPH. Heo et al. (2020) performed stroke prognosis based on the document-level and sentence-level representations of MRI records. Our work extends this line of research by using contextual deep representations of clinical texts to perform survival analysis.

3 Methods

3.1 Task

We first formulate the survival analysis problem. In the discrete context, we divide the continuous time into disjoint intervals $V = (t_{l-1}, t_l]$ where t_0 and t_T are the first and last observation interval boundaries. At time t_u, the model predicts the survival probability in the prediction window $(t_u, t_T]$ with longitudinal features in the observation window $(t_0, t_u]$ (Figure 1).

For each participant i, the survival probability at each time t_l ($l > u$) is $S_i(t_l) = Pr(z > t_l)$, where z is the time-to-event, time until death in our case. The hazard rate of the survival probability is $\lambda_i(t_l) = \frac{S_i(t_{l-1}) - S_i(t_l)}{S_i(t_l)}$.

3.2 Model

Our framework consists of two stages: feature extraction and survival analysis (Figure 1).

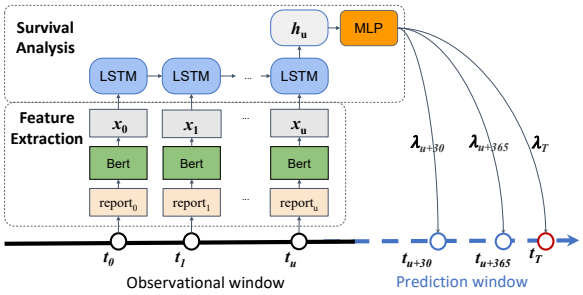

Figure 1: Model Architecture.

3.2.1 Feature Extraction

The input of each time t_l is given by the features extracted from the reports of each patient i. In this work, we evaluate two sets of predefined features and hidden features of the reports.

The first feature set consists of 14 common radiographic findings in computed tomography (CT) imaging reports (aortic aneurysm, ascites, atelectasis, atherosclerosis, cardiomegaly, enlarged liver, gall bladder wall thickening, hernia, hydronephrosis, lymphadenopathy, pleural effusion, pneumonia, previous surgery, and pulmonary edema). The findings are extracted using the convolutional neural network provided by Pandey et al. (2020) which had the reported performance of 0.90 F1 in average.

The second feature set consists of 14 predefined findings in CheXpert (Irvin et al., 2019) which are commonly found in radiology reports (atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, support devices, and normal). These features are extracted using CheXbert (Smit et al., 2020) with the reported performance of 0.80 F1 in average.

For deep representations, we use CheXbert's final hidden layer features of the reports. More

specifically, we extract the information before it passes on to the output layer that consists of 14 linear heads. The representations are vectors of size 768.

Lastly, we construct sequential deep representations by creating arrays of up to three most recent reports of each patient. As the reports can change over time based on the patient's condition, these features are time-variant and contain temporal information that cannot be obtained by a single report. In addition to CheXbert, we apply BERT variations – BERT, BioBert, ClinicalBert and BlueBert.

3.2.2 Deep survival analysis

In this study, the hazard rate has the form

$$\lambda_i(t_l \mid X_i) = \lambda_u(t_l)e^{\psi(X_i)} \quad (1)$$

ψ is a patient's log-risk of failure, X_i are covariates representing a patient's variables up to t_u, and λ_u baseline hazard at t_u.

For the standard Cox Proportional-Hazards (CPH) model (Cox, 1972), $\psi(X_i)$ has the form of a linear combination of p covariates $\beta_1 X_{i1} + \cdots + \beta_p X_{ip}$. In our experiments, the covariates are the features extracted from the reports.

ψ can also be a non-linear risk function of a multilayer perceptron (MLP). To this end, our model is the same as DeepSurv (Katzman et al., 2018).

Both CPH and DeepSurv cannot incorporate the higher-order temporal information from report sequences. To solve this problem, we define $\psi = LSTM(X_i)$ to model the possible time-variant effects of the covariates leading up to t_u (Figure 1). Our model is similar to RNN-SURV (Giunchiglia et al., 2018) and DRSA (Ren et al., 2019). The main difference is that the objective function is the average partial log-likelihood (Kvamme et al., 2019):

$$-\frac{1}{N}\sum_{i\in U_l}\left(\tilde{\psi}(x_i) - \log\sum_{j\in R_l} e^{\tilde{\psi}(x_j)}\right) \quad (2)$$

U_l is the set of patients that are deceased or last known to be alive (censored) by time point t_l. R_l is the set of all live and uncensored patients before t_l. N is the total number of deceased patients in the dataset.

4 Experiments

4.1 Data

The dataset (Pandey et al., 2020) is a collection of thoracoabdominal CT reports in English for heart failure patients from the New York-Presbyterian/Weill Cornell Medical Center who were admitted and discharged with billing codes ICD-9 Code 428 or ICD-10 Code I50 from January 2008 and July 2018 (Table 1). It was reviewed by the institutional board and de-identified. We use each patient's three most recent reports or zero vectors for any missing ones. Their time-to-event is calculated as the number of days between the most recent report date and death date if deceased or the last follow-up date if censored. We perform simple preprocessing steps to confirm each patient has at least one report and nonnegative time-to-event.

Characteristics	n
Number of Patients	11,971
30 days mortality	1,209
365 days mortality	2,062
Total mortality	2,602
Number of Reports	39,752
Avg number of words in reports	306

Table 1: Dataset Overview.

4.2 Metrics

To assess the discriminative accuracies of our models, we use the C-index (Harrell et al., 1982) and time-dependent area-under-the-curve (AUC) (Heagerty and Zheng, 2005), some of the most commonly used evaluation metrics in clinical research (Kamarudin et al., 2017; Pencina and D'Agostino, 2004; Uno et al., 2011). Intuitively, the C-index measures the extent to which the model is able to assign logical risk scores. An individual with shorter time-to-event T should have a higher risk score R than the ones with longer time-to-event. Formally, it is defined as:

$$C = \frac{\sum_{i,j} I(T_i > T_j)\cdot I(R_i < R_j)\cdot d_j}{\sum_{i,j} I(T_i > T_j)\cdot d_j} \quad (3)$$

$$I(c) = \begin{cases} 1 \text{ if } c \text{ is true} \\ 0 \text{ else} \end{cases} \qquad d_j = \begin{cases} 1 \text{ if } T_j \text{ exists} \\ 0 \text{ else} \end{cases}$$

Both C-index and AUC assign a random model 0.5 and a perfect model 1. We measure all-time C-index, C-index at 30 days (C-index@30), and AUC at 30 days and 365 days (AUC@30 and AUC@365) to show the models' performances dealing with different time-to-events[1].

[1] https://github.com/sebp/scikit-survival

Model	C-index	C-index@30	AUC@30	AUC@365
CPH + Feature Set 1	0.499 ± 0.025	0.504 ± 0.032	0.503 ± 0.037	0.501 ± 0.028
CPH + Feature Set 2	0.621 ± 0.014	0.632 ± 0.033	0.642 ± 0.030	0.642 ± 0.030
CPH + Hidden Features	0.674 ± 0.023	0.696 ± 0.022	0.710 ± 0.022	0.697 ± 0.026
MLP + Feature Set 1	0.502 ± 0.023	0.509 ± 0.026	0.509 ± 0.030	0.501 ± 0.032
MLP + Feature Set 2	0.658 ± 0.010	0.671 ± 0.025	0.685 ± 0.023	0.683 ± 0.008
MLP + Hidden Features	0.704 ± 0.017	0.726 ± 0.020	0.744 ± 0.019	0.734 ± 0.018
LSTM + Sequential HF	$\mathbf{0.709 \pm 0.022}$	$\mathbf{0.733 \pm 0.031}$	$\mathbf{0.752 \pm 0.033}$	$\mathbf{0.742 \pm 0.023}$

Table 2: Evaluation results. Feature Set 1 - (Pandey et al., 2020), Feature Set 2 - (Irvin et al., 2019), Sequential HF - sequential hidden features

Model	C-index	C-index@30	AUC@30	AUC@365
BERT-Base (Devlin et al., 2019)	0.603 ± 0.115	0.611 ± 0.123	0.618 ± 0.134	0.620 ± 0.136
BioBert (Lee et al., 2020)	0.701 ± 0.021	0.714 ± 0.027	0.734 ± 0.029	0.739 ± 0.028
ClinicalBert (Alsentzer et al., 2019)	0.692 ± 0.019	0.705 ± 0.023	0.723 ± 0.025	0.727 ± 0.029
BlueBert (Peng et al., 2019)	$\mathbf{0.713 \pm 0.019}$	$\mathbf{0.735 \pm 0.024}$	$\mathbf{0.755 \pm 0.024}$	$\mathbf{0.756 \pm 0.021}$
CheXbert (Smit et al., 2020)	0.709 ± 0.022	0.733 ± 0.031	0.752 ± 0.033	0.742 ± 0.023

Table 3: Evaluation results of LSTM + sequential hidden features using different BERT models.

4.3 Training

We perform a grid search to find the optimal hyper-parameters based on the metrics and use them for all configurations. The learning rate is set to 0.0001 with an Adam optimizer. We iterate the training process for 100 epochs with batch size 256 and early stop if the validation loss does not decrease. The dropout rate is 0.6. We perform five-fold cross-validation to produce 95% confidence intervals for each metric. The training, validation and test splits are 70%, 10%, 20%, respectively. We use pycox and PyTorch to implement the framework[2]. The end-to-end training takes about 30 minutes with NVIDIA Tesla P100 16 GB GPU, mainly due to feature extraction.

4.4 Results & Discussions

Table 2 shows our experimental results with variations in covariates and survival analysis models.

Our LSTM model with hidden features (LSTM + Hidden Features) achieves the best results (0.709 in C-index), 3.5% and 0.5% improvements over CPH + Hidden Features and MLP + Hidden Features. In contrast to the MLP, its data included the reports from patients' prior visits with more textual and higher-order temporal information. Nonetheless, the improvements are stll marginal, suggesting that the evaluation of the effectiveness of LSTM in survival analysis in this context would require more empirical evidence, particularly with more longitudinal text data.

We observe that the hidden features provide at least 5% improvements over the other feature sets with both CPH and MLP. This indicates that the hidden features capture textual information more thoroughly than the predefined features for survival analysis.

We find that Feature Set 2, obtained with CheXbert (Smit et al., 2020), performs significantly better ($> 10\%$ C-index) than Feature Set 1, obtained with the CNN model (Pandey et al., 2020). With both CPH and MLP, Feature Set 1 yields around 0.5 in C-index and AUC, whereas Feature Set 2 shows prognostic value in the 0.62-0.69 range. The difference of the feature sets directly results in the performance difference. While Feature Set 1 and Feature Set 2 have overlapping features (atelectasis, cardiomegaly, pleural effusion, and pneumonia), Feature Set 1 is not as discriminatory as Feature Set 2. This observation informs us that much important textual information with prognostic value is likely lost between the feature sets.

Finally, we compare our model on BERT-Base variants. BERT-Base, CheXbert and BlueBert used "uncased" text. BioBert and ClinicalBert used "cased" text. BioBert was pretrained on PubMed abstracts. ClinicalBert was initialized with BioBert's weights and further trained on MIMIC-III clinical

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

notes. BlueBert was pretrained on both datasets altogether. Table 3 shows that all BERT variants (except the original BERT) capture the textual information more comprehensively than the predefined features and yield significantly more accurate predictions. Further, the models with more pertinence to radiology reports perform incrementally better. BlueBert outperforms others and improves on CheXbert slightly. This observation is consistent with the findings in (Peng et al., 2019). These results illustrate that using hidden layer representations in survival analysis can generalize across deep learning models based on their areas of focus.

4.5 Conclusion & Future Work

Incorporating the textual information of clinical texts in survival analysis is challenging because of their unstructured format. Automatic extraction tools have a small set of features selected by experts and fail to capture the information fully and precisely. We show a novel method of using hidden layer representations of clinical texts as covariates for proportional hazards models. When applied to predicting all-cause mortality of heart failure patients, the results indicate that hidden features encapsulate more comprehensive and effective textual information than predefined features.

We plan to explore the use of the attention mechanism to the input sequence and test the generalizability of this method with more datasets. In addition, we plan to gain more insights on how hidden features are influenced (e.g. word choice, text length, etc.) and add value for better prediction as interpretability is highly important in the medical domain. We hope our small contribution provides assistance in the scalable development of accurate predictive models that harness clinical text information.

Acknowledgment

Research reported in this publication was supported by National Library of Medicine - National Institutes of Health (NIH) under award number R00LM013001. It was also supported by National Heart, Lung, and Blood Institute of NIH under award number R01HL1276610. This study also received support from NewYork-Presbyterian Hospital (NYPH) and Weill Cornell Medical College (WCMC), including the Clinical and Translational Science Center (CTSC) (UL1TR002384) and Joint Clinical Trials Office (JCTO). Additionally, we would like to thank Dr. Pranav Rajpurkar and Dr. Matthew P Lungren for providing the ChexBert model.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Wendy Chapman, Will Bridewell, Paul Hanbury, Gregory Cooper, and Bruce Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310.

D. R. Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Eleonora Giunchiglia, Anton Nemchenko, and Mihaela van der Schaar. 2018. RNN-SURV: a deep recurrent model for survival analysis. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, volume 11141, pages 23–32.

Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. 1982. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546.

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

Tak-Sung Heo, Chulho Kim, Jeong-Myeong Choi, Yeong-Seok Jeong, and Yu-Seop Kim. 2020. Various approaches for predicting stroke prognosis using magnetic resonance imaging text records. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 1–6.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. 2017. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):53.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1).

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30.

Curtis P. Langlotz. 2015. *The radiology report: a guide to thoughtful communication for radiologists and other medical professionals*. Curtis P. Langlotz.

Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. 2017. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific Reports*, 7(1):10353.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36(4):1234–1240.

Matthew B.A. McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. 2020. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 913–927.

Annelaura B Nielsen, Hans-Christian Thorsen-Meyer, Kirstine Belling, Anna P Nielsen, Cecilia E Thomas, Piotr J Chmura, Mette Lademann, Pope L Moseley, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Anders Perner, and Søren Brunak. 2019. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *The Lancet Digital Health*, 1(2):e78–e89.

Mohit Pandey, Zhuoran Xu, Evan Sholle, Gabriel Maliakal, Gurpreet Singh, Zahra Fatima, Daria Larine, Benjamin C. Lee, Jing Wang, Alexander R. van Rosendael, Lohendran Baskaran, Leslee J. Shaw, James K. Min, and Subhi J. Al'Aref. 2020. Extraction of radiographic findings from unstructured thoracoabdominal computed tomography reports using convolutional neural network based natural language processing. *PLOS ONE*, 15(7):1–15.

Michael J Pencina and Ralph B D'Agostino. 2004. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*, 23(13):2109—2123.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. In *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, volume 2017, pages 188–196.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 58–65.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. 2016. Deep survival analysis. *arXiv:1608.02158 [cs, stat]*.

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. 2019. Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4798–4805.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Charles L. Sprung, Gavin M. Joynt, Michael D. Christian, Robert D. Truog, Jordi Rello, and Joseph L. Nates. 2020. Adult ICU Triage During the Coronavirus Disease 2019 Pandemic: Who Will Live and Who Will Die? Recommendations to Improve Survival*. *Critical Care Medicine*, 48(8):1196–1202.

Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. 2011. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117.