Mitigating Membership Inference in Deep Survival Analyses with Differential Privacy

Liyue Fan
Dept. of Computer Science
University of North Carolina at Charlotte
Charlotte, NC
liyue.fan@uncc.edu

Luca Bonomi
Dept. of Biomedical Informatics
Vanderbilt University Medical Center
Nashville, TN
luca.bonomi@vumc.org

***Abstract*—Deep neural networks have been increasingly integrated in healthcare applications to enable accurate predicative analyses. Sharing trained deep models not only facilitates knowledge integration in collaborative research efforts but also enables equitable access to computational intelligence. However, recent studies have shown that an adversary may leverage a shared model to learn the participation of a target individual in the training set. In this work, we investigate privacy-protecting model sharing for survival studies. Specifically, we pose three research questions. (1) Do deep survival models leak membership information? (2) How effective is differential privacy in defending against membership inference in deep survival analyses? (3) Are there other effects of differential privacy on deep survival analyses? Our study assesses the membership leakage in emerging deep survival models and develops differentially private training procedures to provide rigorous privacy protection. The experimental results show that deep survival models leak membership information and our approach effectively reduces membership inference risks. The results also show that differential privacy introduces a limited performance loss, and may improve the model robustness in the presence of noisy data, compared to non-private models.**

***Index Terms*—Deep Learning, Survival Analysis, Membership Inference, Data Privacy**

I. Introduction

Deep learning has been increasingly applied in healthcare research and applications [1]. Deep models are shown to learn representations of health data effectively using multiple levels of abstraction, thus enabling accurate predicative analyses. Training these models often requires large datasets and sophisticated computational infrastructure. Therefore, researchers who do not have access to adequate data and computational resources would rely on pre-trained models shared by others. In addition, model sharing also benefits collaborative research efforts [2], facilitating knowledge transfer and integration among multiple institutions. However, explicit privacy concerns may rise regarding what information these models may reveal about individual data contributors (i.e., patients in the training set). In fact, recent studies have shown that by accessing a shared model, an adversary may be able to learn the participation of a target individual in the training set and even infer sensitive attribute values [3], [4].

To facilitate the application of deep learning techniques in healthcare, it is imperative to develop privacy-protecting solutions to safeguard the data used for training deep models. Differential privacy [5], a rigorous and provable privacy notion, has become the state-of-the-art paradigm for protecting sensitive data. In a nutshell, differential privacy ensures that an adversary would not be able infer whether an individual is in the data, by observing computational results. Recently, differential privacy has been extended to training deep models, protecting the underlying training samples [6]. The framework allows a data curator to train deep learning models using the classic stochastic gradient descent method while satisfying differential privacy. In this work, we investigate the applicability of differentially private deep learning in healthcare, its efficacy in defending against privacy attacks on deep models, and its usability in clinical predictive tasks.

We focus on survival analysis, where a number of deep learning approaches [7]–[10] have been recently proposed to improve upon traditional methods (such as the Cox proportional hazards model [11]) with neural networks. In contrast to prior research, which has studied the privacy risks and solutions associated with Kaplan–Meier time-to-event analyses [12], [13], this work investigates new privacy challenges in incorporating patient covariates and deep neural network models. The contribution of our work is three-fold. For deep survival models, we develop a quantifiable measure of membership leakage via membership inference attacks, which takes into account the input covariates and the predicted survival functions. Empirically, we show that deep survival models leak membership information, potentially disclosing individual participation in the training set. Furthermore, we develop differentially private training procedures for those survival models based on the private framework [6], and evaluate the efficacy in defending against membership inference. Moreover, we empirically study the impact of differential privacy on deep survival models, in terms of convergence, performance, and robustness. Our analyses show that rigorous privacy can be achieved for deep survival analysis models and membership inference can be mitigated. While privacy may impact the predictive accuracy, the accuracy loss is moderate; private models may be more resilient to noise in the data compared to non-private models.

The rest of the paper is organized as follows. Section II introduces recently proposed deep learning approaches to

survival analyses; Section III describes the proposed membership inference attacks on deep survival models; Section IV introduces the background on differential privacy and the training procedure for differentially private deep survival models; Section V presents the empirical evaluation methodology and discusses the results; Section VI concludes the paper and states future work opportunities.

II. Deep Learning for Survival Analysis

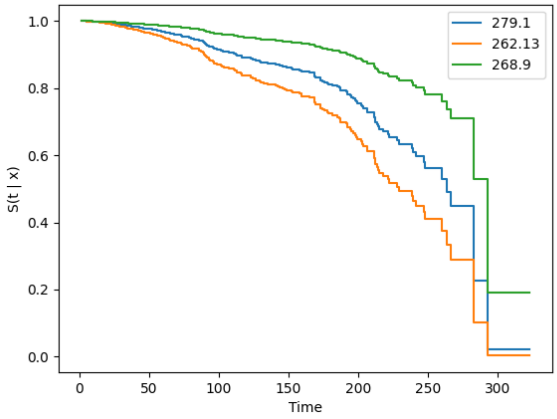
In a survival study, a dataset $D = \{(x^i, t^i, e^i)\}_{i=1}^N$ is often given which contains N observed instances/patients. For patient i, x^i is the covariate vector; t^i is the time when the event or censoring occurred; e^i is the event or censoring that occurred at t^i. Note that t^i is either the time when the event (e.g., death) occurred or the time when the patient was censored (e.g., not following up); in either case, the patient was known to be alive prior to time t^i. Survival models are trained with the labeled dataset D to optimize specific objectives, e.g., Cox partial likelihood in the Cox proportional hazards model.

Recent works have shown that deep neural network models can outperform traditional survival methods (e.g., the Cox proportional hazards model), by modeling both linear and nonlinear effects from covariates. In this work, we consider four emerging deep learning approaches to survival analyses, namely: DeepSurv [7], DeepHit [8], Nnet [9], and CoxTime [10]. DeepSurv adapted the Cox proportional hazards model to neural networks and showed that novel networks were able to outperform classic Cox models. Given larger datasets and higher computational power, recent studies proposed to study non-proportional hazards by adopting a time-dependent risk function (i.e., CoxTime) or discrete-time models (i.e., DeepHit and Nnet).

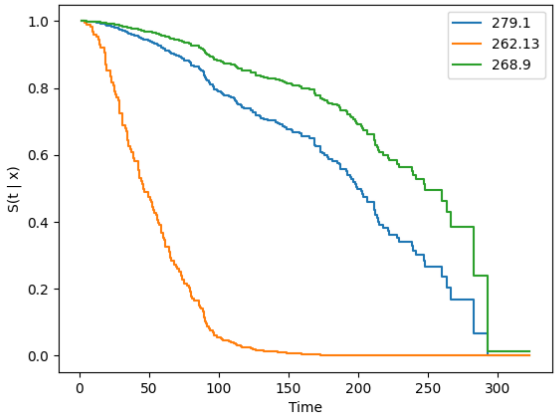
At inference time, a deep survival model takes as input patient j's covariates and predicts the survival function denoted as $S(t|x^j)$, which is the probability that patient j survived beyond time t, for any t in the time span of the study. The survival function $S(t|x^j)$ is used to assess the performance of a deep survival model, e.g., in time-dependent concordance. It is also used in our proposed membership inference attacks to quantify the privacy leakage.

Although different deep learning approaches may vary in the model's output, it is not difficult to derive the survival function from the output of any deep survival model studied in this work. For example, DeepHit directly estimates the conditional probability for the event distribution and Nnet directly estimates the conditional probability of surviving each interval; both DeepSurv and CoxTime estimate the log-risk function as in the Cox model, and adopt the Breslow estimator to compute the baseline hazard. We refer the readers to original manuscripts [7]–[10] for additional model details.

III. Membership Inference on Deep Survival Models



(a) Including those individuals in training



(b) Excluding those individuals

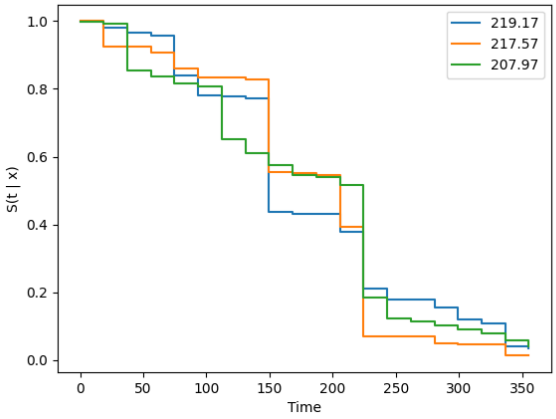
Fig. 1: DeepSurv Predicted Survival Functions for Sample Individuals in METABRIC (best viewed in color): survival functions may differ for the same individuals depending on whether they are included in the training set; legend indicates ground truth time-to-event.

Recent research has shown that deep models are prone to overfitting and thus may leak sensitive information about training data [3], [4]. In this work, we investigate the leakage of *membership* information, which discloses an individual's participation in the training set.

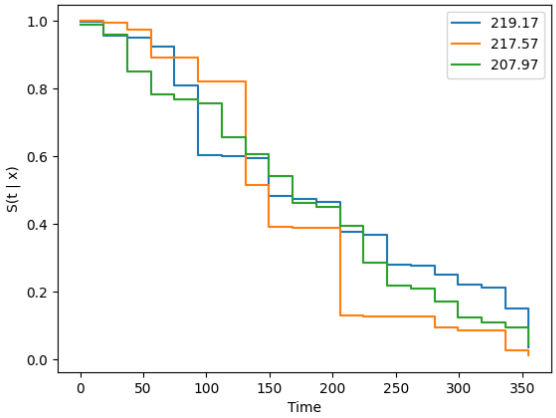
A. Membership Leakage of Deep Survival Models

In a well-known study [3], researchers quantified the membership leakage of deep models in the black-box setting, where an adversary's access to the target model is limited to the model's output on a given input (i.e., query). The target model in the study [3] is assumed to be a classifier, which outputs a vector where each element indicates the probability for the corresponding class. While adopting the same adversarial assumptions, this work focuses on deep survival models as target models, for which the privacy risks are not well understood.

Our hypothesis is that the predictions of deep survival models, specifically survival functions, may leak member-



(a) Including those individuals in training



(b) Excluding those individuals

Fig. 2: DeepHit Predicted Survival Functions for Sample Individuals in METABRIC (best viewed in color): survival functions may differ for the same individuals depending on whether they are included in the training set; legend indicates ground truth time-to-event; note that DeepHit is a discrete-time model and survival functions are estimated over 20 time intervals.

ship information. As an illustration, two qualitative studies with continuous-time (e.g., DeepSurv) and discrete-time (e.g., DeepHit) models are presented in Figure 1 and Figure 2. Specifically, in each study, we trained two models with the METABRIC dataset (see Section V for dataset information): model (**a**) trained with 80% of the dataset and model (**b**) with the same training set but excluding three random individuals[1]. In each figure, we plot the predicted survival functions by model (a) and model (b) for those three individuals. It can be observed that the presence/absence of those individuals in the training set has a noticeable impact on the predictions. As can be seen in Figure 1, the survival probabilities predicted by model (a) drop significantly for all individuals around their true time-to-events; on the other hand, model (b) predicts only

[1]For visualization, we randomly selected three individuals with similar time-to-events.

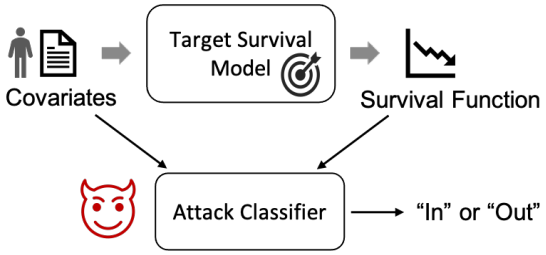


Fig. 3: Membership Inference Attacks on Deep Survival Models.

moderate decreases for two individuals and near zero survival probability much earlier for the third individual (i.e., with true time-to-event at 262.13). Similarly in Figure 2, model (b) predicts moderate survival probability decreases for the blue and green individuals around their true time-to-events and an under-estimated survival probability for the orange individual, before the actual time-to-event. Those results illustrate that for individuals in the training set, a model is more likely to predict survival probabilities that resemble their true time-to-events. In the next subsection, we will devise an attack model based on our observation to quantify such membership leakage. In Section V, we will present empirical evidence that deep survival models leak membership information.

B. Membership Inference Attacks on Deep Survival Models

To quantify the membership leakage in deep survival models, we adapted the membership inference attack proposed in [3] to our setting. A basic assumption is the adversary has prior knowledge of the covariates of a patient and has black-box access to the target model. The goal of the adversary is to infer whether the patient participated in the training set. It is also assumed that the adversary can sample records from the training distribution (although non-overlapping with the secrete training set) and can train shadow models to simulate the behaviors of the target model.

The attack framework is depicted in Figure 3. First the adversary queries the target survival model with the covariate vector of an individual and obtains the predicted survival function. Then, the adversary feeds the covariates and the survival function to an attack model, which predicts whether the individual was used to train the target model. As can be seen, the attack model is a binary classifier, predicting "in" or "out" labels for each individual. In principle, any binary classifier could be used by the adversary; in our work we adopted a fully connected neural network with one hidden layer of 64 nodes with ReLU activations and a softmax layer, as suggested in [3]. To train the attack classifier, we used 20 shadow models to simulate the behaviors of the target model. Note that our shadow models are also deep survival models with the same architecture as the target model. Following the suggestion in [3], we trained shadow models using data sampled from the same populations but disjoint from the target model's training set.

To consistently evaluate the outputs of all deep survival models, we provide the attack classifier with survival functions predicted over 20 discrete intervals (as defined for discrete-time models DeepHit and Nnet); for continuous-time survival

models (i.e., DeepSurv and CoxTime), we post-process their output survival functions to estimate the survival probabilities over those intervals accordingly.

IV. Learning Deep Survival Models with Differential Privacy

The privacy model adopted in our work is differential privacy (DP) [5]. In recent years, differential privacy has become the state-of-the-art privacy paradigm for protecting statistical databases, as it provides provable privacy protection. Relevant to our work, recent research has shown that DP provides potential defense against membership inference attacks in machine learning applications [3].

A. Differential Privacy Background

Intuitively, an algorithm A satisfying DP ensures that an adversary, who observes the output of A, cannot determine whether any particular individual record was included in the input. The DP notion aims at achieving indistinguishably for any pair of neighboring databases D, D', which differ in at most a single record (i.e., data of an individual), thus protecting the presence of the record. Specifically, a randomized algorithm A is (ϵ, δ)-differentially private if for any two neighboring databases D, D' and any subset $S \in Range(A)$:

$$\Pr[A(D) \in S] \leq \exp(\epsilon) \Pr[A(D^{'}) \in S] + \delta. \quad (1)$$

The privacy parameter $\epsilon > 0$ bounds the difference between output probabilities of neighboring databases. ϵ is often referred to as the privacy budget. The parameter $\delta \in [0, 1]$ accounts for the probability of a privacy breach. In practice, ϵ, δ parameters help data curators control the information leakage. Typically, smaller ϵ and δ values indicate stronger privacy protection and possibly lower accuracy, and vice versa.

B. Training Deep Survival Models

In this work, we aim at applying DP to training deep survival models, in order to share those models while protecting the sensitive training data. DP guarantees in deep learning applications can be achieved by clipping and perturbing the gradient during training [6], which ultimately limits the overall influence of any individual training example on the model. To account for differential privacy across training epochs, the moments accountant approach has been proposed [6], which provides stronger estimates of privacy loss compared to other composition theorems [5].

We modify the training procedures for deep survival models to satisfy (ϵ,δ)-DP. Specifically, we clip the ℓ_2 norm of each per-sample gradient with a threshold C and perturb the average gradient in a batch with a Gaussian noise draw from $\mathcal{N}(0, \sigma^2 C^2)$ to protect each training sample. We adopt the Rényi Differential Privacy (RDP) Accountant [14] to track the differential privacy budget ϵ spent in training. In fact, ϵ can be effected by a number of factors, including the input data size, batchsize, σ, and the number of training epochs. We include below a snippet of the Python code that illustrates the modified training procedure implemented with PyTorch and Opacus libraries.

```
## initialize RDP accountant,
## DP optimizer, and survival model
accountant = RDPAccountant()
dp_optimizer = DPOptimizer(
    optimizer=optimizer,
    noise_multiplier=sigma,
    max_grad_norm=C,
    expected_batch_size=batch_size)

dp_optimizer.attach_step_hook(
    accountant.get_optimizer_hook_fn(
    sample_rate=batch_size/len(x_train)))

model = DeepSurv(net, dp_optimizer)

...model fitting...

## get total privacy loss
epsilon, alpha =
    accountant.get_privacy_spent(
    delta=1/len(x_train))
```

It is important to choose the privacy parameters to strike a balance between privacy and model utility. In private deep learning, higher ϵ values are often considered, e.g., $\epsilon = 8$ as in [6]. In our study, we consider $\epsilon \leq 16$ to provide privacy protection in deep survival analyses without incurring high utility loss. We set $\delta = \frac{1}{|D|}$ as recommended in [5], where $|D|$ represents the size of the training set, in order to protect every individual in the dataset.

V. Empirical Results

A. Evaluation Methodology

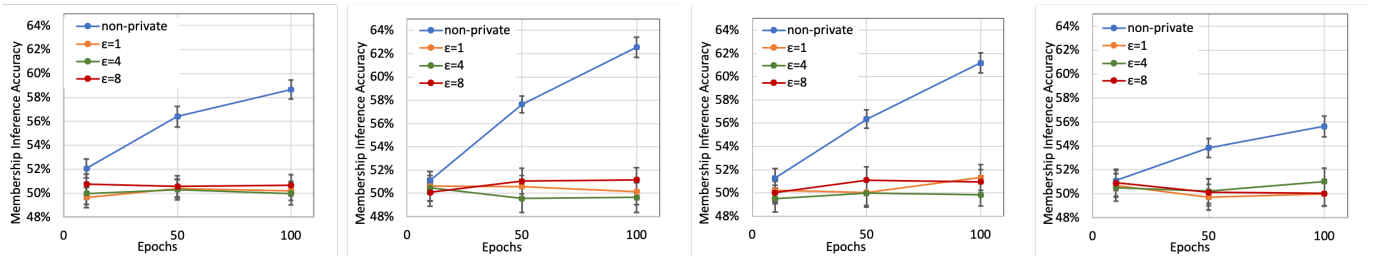
Data Description. We adopt four real-world survival datasets in the evaluation. Three datasets were used in DeepSurv [7]: the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), the Rotterdam tumor bank and German Breast Cancer Study Group (GBSG), and the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT). In addition, we include NWTCO from the National Wilm's Tumor Study [15] obtained through RDatasets[2]. Table I provides a summary of the datasets.

TABLE I: Summary of Data

Name	Size	Covariates	Time-to-Events	Perc. Censored
METABRIC	1904	9	0 to 355.2 months	42%
GBSG	2232	7	0.3 to 87.4 months	43%
SUPPORT	8873	14	3 to 2029 days	32%
NWTCO	4028	6	4 to 6209 days	86%

Data Processing. For all datasets, we standardize the numerical covariates and encode the categorical covariates with entity embedding as in [10]. Time-to-events are discretized

[2]https://vincentarelbundock.github.io/Rdatasets/



(a) DeepSurv (b) DeepHit (c) Nnet (d) CoxTime

Fig. 4: Membership Inference for Non-Private and Private Models with METABRIC (best viewed in color): models were trained over the specified number of epochs; in addition, private models were trained to meet the target ϵ values.

into 20 equidistant intervals for DeepHit and Nnet. As time-to-event is regarded as a regular covariate by CoxTime, it is also standardized for training CoxTime models.

Utility Measures. In survival analysis, the concordance index (C-Index) [16] is commonly applied to evaluate the *discrimination* performance of survival models. We adopt the time-dependent concordance (C^{td}) definition [17] which utilizes the whole predicted survival function: a subject who developed the event should have a less predicted probability of surviving beyond his/her survival time than any subject who survived longer. Formally, the C-index is computed by

$$C^{td} = \Pr[S(t^i|x^i) < S(t^i|x^j) \,|\, t^i < t^j, e^i = 1]. \quad (2)$$

To evaluate model *calibration* performance, we adopt the integrated Brier score (IBS), which is based on the Brier score for censored data [18]. Specifically, we compute the following

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{S(t|x^i)^2 \mathbb{1}\{t^i < t, e^i = 1\}}{G(t^i)} + \frac{(1 - S(t|x^i))^2 \mathbb{1}\{t^i > t\}}{G(t)}\Big] \quad (3)$$

$$IBS = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} BS(s)\, ds \quad (4)$$

where $G(t)$ is the Kaplan-Meier estimate of the censoring survival function. In our experiments, we adopt the numerical integration for IBS and break down the time span for each dataset to 20 intervals, in order to be consistent with the discretization of time-to-events for DeepHit and Nnet.

Model Implementation and Hyperparameters. We implement deep survival models in PyTorch and differentially private training with Opacus[3]. The networks are standard multi-layer perceptrons with 2 layers and 32 nodes in each layer, ReLU activation, and dropout=0.1. Batchsize is set to 256 and the number of training epochs is varied from 10 to 100. We adopt the Adam optimizer with 0.01 learning rate. For DeepHit, parameter α specifies the weight of the ranking loss term in the total loss. We set $\alpha = 4$ to achieve a balance between discrimination performance and calibration performance. For training private models, the parameter σ is set to meet the specified ϵ value and the clip norm C is set to 32 (for DeepHit and CoxTime) and 4 (for DeepHit and Nnet) for better convergence.

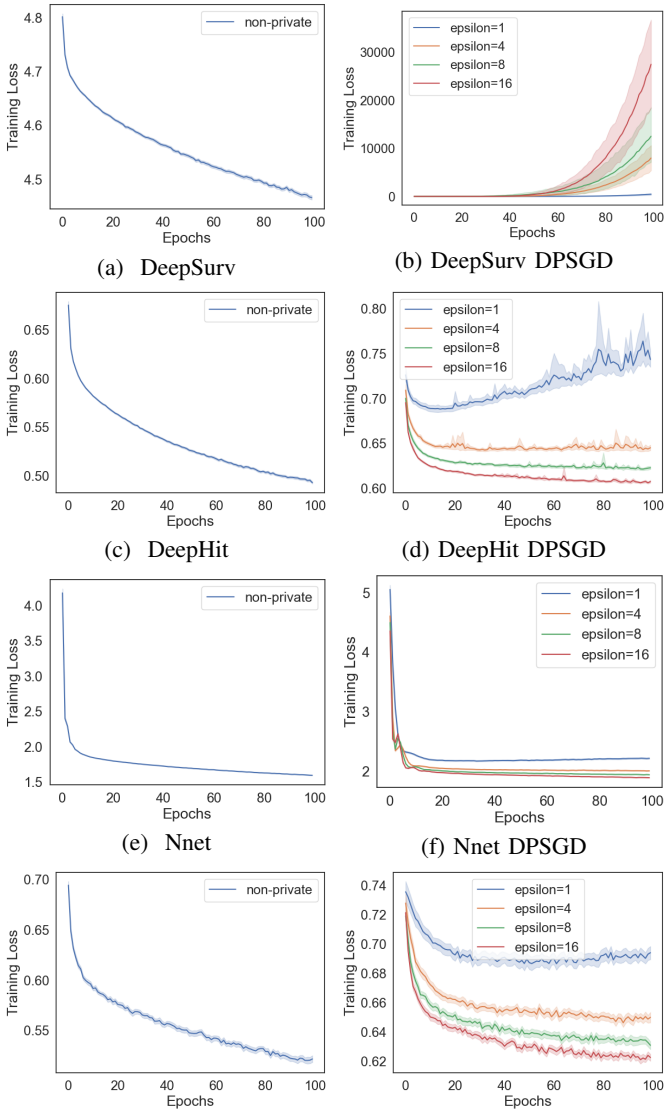
[3]https://github.com/pytorch/opacus

B. *Efficacy against Membership Inference*

First, we examine the amount of membership leakage in deep survival models and whether differential privacy can help mitigate such privacy risks. Specifically, we are interested in the effects of the number of training epochs. Intuitively, with a higher number of epochs, the model may "memorize" more of each training sample, hence higher membership leakage.

In Figure 4, we plot the accuracy of membership inference attacks against deep survival models trained in a number of privacy settings, i.e., non-private and $\epsilon \in [1, 4, 8]$. Each experiment was run 50 times, and we reported the mean testing accuracy and 95% confidence interval. As a binary classification problem, an accuracy above 50% indicates positive leakage of membership information. When trained over 10 epochs, both private and non-private models do not inflict significant membership leakage. Private models at different ϵ values yield around 50% membership inference accuracy, while the accuracy of non-private models is slightly higher than 50%. As the number of training epochs increases from 10 to 100, we observe a steady increase of accuracy among all non-private models, rising above 60% for DeepHit and Nnet. However, for private models, the membership inference accuracy remains around 50%, indicating effective defense against such attacks. Note that deep survival models may be trained over a very large number of epochs in non-private settings, e.g., 1000 epochs as in Nnet [9]. Our results show that we must be mindful of additional membership leakage that may be inflicted by training the model over more epochs.

While in theory higher ϵ values provide weaker differential privacy guarantees, we observe only a marginal increase of membership inference accuracy for setting $\epsilon = 8$, compared to models that satisfy $\epsilon = 1$ and 4. It is likely that the clipping and perturbation operations required by differential privacy introduce uncertainty to the training process, even at higher ϵ settings. In the following, we will further examine the impacts of differential privacy on training deep survival models, in terms of training stability and model utility.



(a) DeepSurv (b) DeepSurv DPSGD

(c) DeepHit (d) DeepHit DPSGD

(e) Nnet (f) Nnet DPSGD

(g) CoxTime (h) CoxTime DPSGD

Fig. 5: Training Loss of Deep Survival Models with METABRIC (best viewed in color): private models were trained to meet the target ϵ values.

C. Convergence

To understand the impact of differential privacy on model training, we examine the model convergence for private and non-private survival models. As deep survival models have custom objectives, below we focus on the comparative analysis between private and non-private models within the same approach. In Figure 5, we plot the model training loss recorded over 100 epochs. For each model, we report the mean and 95% confidence interval among 100 runs. As can be seen, all non-private models converge nicely, i.e., exhibiting reduced loss as training proceeds. Furthermore, their convergence does not deviate significantly among 100 independent runs. We notice that Nnet converges faster than other models, thanks to its discrete-time loss function which allows for rapid training with mini-batches.

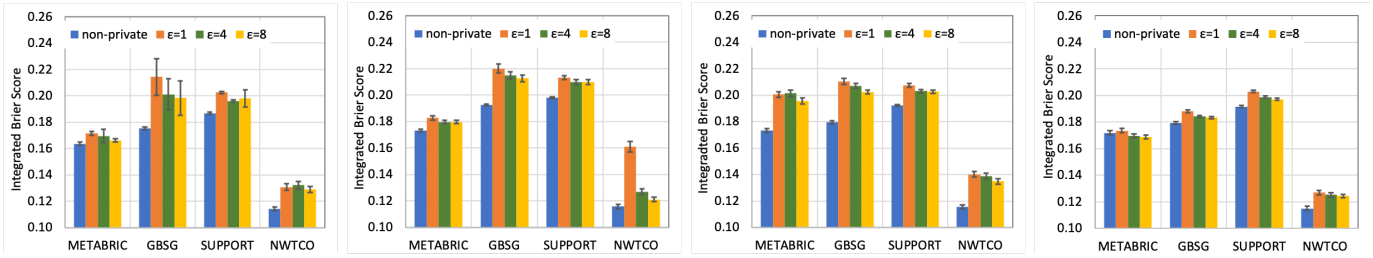
On the other hand, we observe several convergence characteristics in private models. Firstly, private models may converge to noisy solutions, i.e., resulting in higher training losses than their non-private counterparts. Secondly, while reducing the training loss initially, private models may not converge in the long run, i.e., exhibiting an increased loss as training proceeds. That can be observed in private DeepSurv models as well as DeepHit and CoxTime models with $\epsilon = 1$. Thirdly, the convergence behavior of private models varies more considerably among independent runs than that of the non-private models, as shown by wider confidence intervals. We believe that clipping and perturbation during private training may change the optimization process, thus leading to varied, noisy solutions or non-convergence. With stronger privacy guarantees, i.e., lower ϵ values, the training process is likely to be perturbed with larger noises, thus exacerbating the instability in training. To learn meaningful models, we set the number of training epochs for private models to 10 and for non-private models to 100, unless specified otherwise.

D. Impacts on Utility

Figure 6 and Figure 7 report our utility evaluation on non-private and private deep survival models, using the integrated Brier score (IBS) and time-dependent concordance index (C-Index) as measures. Mean and 95% confidence interval were reported among 100 runs.

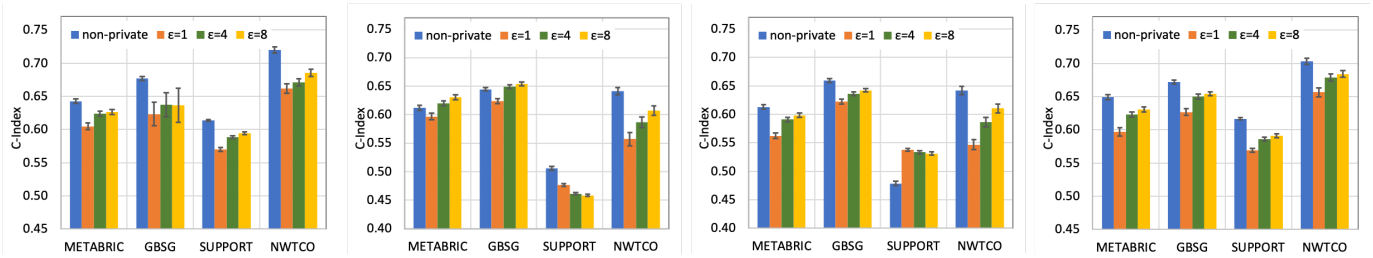
For IBS, lower scores indicate higher calibration performance. The results shown in Figure 6 are not surprising. Non-private survival models achieve the lowest IBS scores for each dataset; among private models, those with weaker privacy guarantees (i.e., higher ϵ values) yield lower IBS scores. From these results, we can observe the negative effects of privacy on model utility. For instance, the performance gap between private and non-private models may be significant, especially for Nnet, DeepHit, and DeepSurv. Private models may yield higher uncertainty (i.e., wider confidence intervals) in their performance, e.g., in the results of DeepSurv models. For C-Index, higher scores indicate higher discriminative performance. The results in Figure 7 show that non-private models often lead to highest C-Index scores. Among private models, increasing ϵ value yields higher C-Index, demonstrating the trade-off between privacy and utility. We note that non-private CoxTime models yield similar IBS scores to those of private counterparts, e.g., for METABRIC, and DeepHit and Nnet private models may produce higher C-Index scores than non-private models. It is possible that using the same architecture and hyper-parameters for all models and datasets may lead to suboptimal performance for non-private models.

The combination of IBS and C-Index results provides a comprehensive performance evaluation of deep survival models. It is can be seen that all survival models perform well with the NWTCO dataset, signified by lower IBS scores and higher C-Index scores; on the other hand, they perform poorly with the SUPPORT dataset, i.e., leading to higher IBS scores and lower C-Index scores. We believe that the neural network can have a custom design for each dataset to optimize



(a) DeepSurv (b) DeepHit (c) Nnet (d) CoxTime

Fig. 6: Integrated Brier Score for Non-Private and Private Models (best viewed in color).



(a) DeepSurv (b) DeepHit (c) Nnet (d) CoxTime

Fig. 7: Concordance Index for Non-Private and Private Models (best viewed in color).

its performance. Furthermore, results for both measures are highly consistent: where a model performs well in IBS, it is likely to perform well in C-Index. However, exceptions can be observed for private DeepHit and Nnet models with the SUPPORT dataset, where some models yield lower IBS scores but also lower C-Index scores, e.g., $\epsilon = 8$ (compared to other private models) and non-private Nnet. We attribute those exceptions to the low performance of the non-private base models.

E. Qualitative: Survival Curves

Although the performance of private deep survival models has been assessed with widely adopted IBS and C-Index measures, we provide a qualitative evaluation in Figure 8 to allow readers to view the predictive outcomes of the studied models. Specifically, we report in Figure 8 the predicted survival functions by non-private and private survival models for sample individuals in the METABRIC dataset.

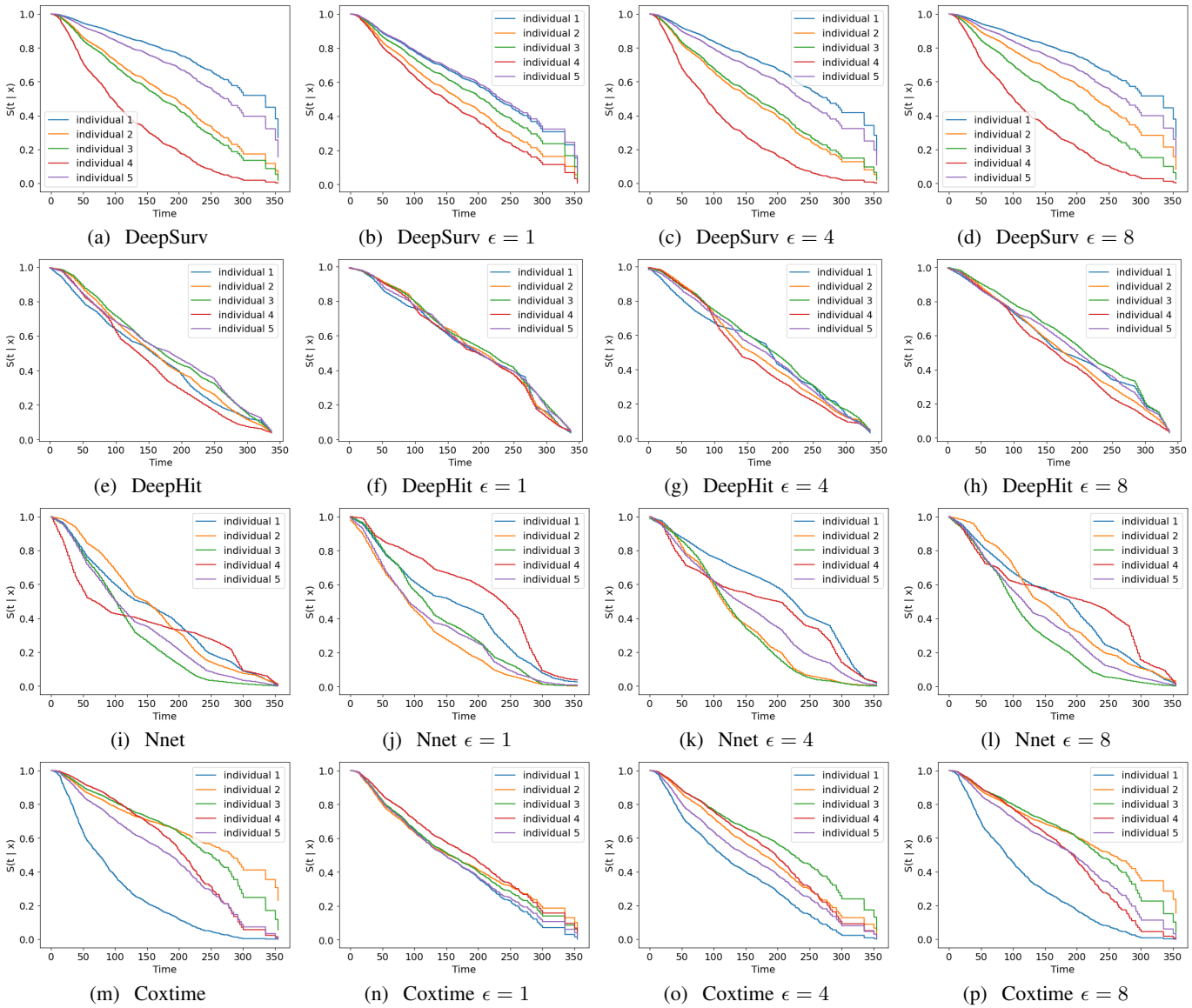
For each approach (i.e., within each row in Figure 8), we trained both non-private and private models using the same training partition and survival functions were predicted for the same set of testing individuals. For discrete time models, i.e., DeepHit and Nnet, we plot the interpolated survival probabilities within each time interval for a smoother visualization. As can be seen, within each approach, the survival functions predicted by private models differ from those predicted by non-private models, for the same test individuals. By relaxing the privacy guarantees, we may obtain more similar survival functions from the private models to those of non-private models. For example, the survival functions predicted by models with $\epsilon = 8$ are most similar to the non-private models' predictions.

F. Case Study: Robustness

Recall that our goal is to enable model sharing such that researchers and collaborators could apply the shared model and perform inference on their local data. However, it may be challenging to maintain model utility at inference time, as local data may come from different distributions compared to that of the training data and/or exhibit quality issues (e.g., containing errors). To that end, we conducted a case study to evaluate the robustness of the shared model. We intentionally introduced controlled noise in testing data, to simulate unanticipated data distributions or quality issues. The utility of deep survival models on the noisy testing set is reported in Figure 9 and Figure 10.

For each deep survival model, we trained both non-private and private models with the same training set and evaluated their performances on the same noisy testing set. Specifically, given a noise parameter $l \in (0, 1]$, for each attribute x_j^i of testing sample x^i, we flipped x_j^i with probability l if x_j^i is binary, and we replaced x_j^i by sampling from the uniform distribution $[(1-l)x_j^i, (1+l)x_j^i]$ if x_j^i is numerical or categorical. We adopted various l values in our case study and larger l values indicate noisier testing data. Each experiment was run 100 times and the mean and 95% confidence interval were reported.

From Figure 9 and Figure 10, we observed lower model performance when increasing l values, indicated by higher IBS scores and lower C-Index scores. This illustrates that both non-private and private models tend to perform poorly if testing data does not come from the same distributions as training data. Furthermore, we observed that private models may outperform non-private models, especially with noisier testing data (i.e., larger l values), resulting in lower IBS scores



(a) DeepSurv (b) DeepSurv $\epsilon = 1$ (c) DeepSurv $\epsilon = 4$ (d) DeepSurv $\epsilon = 8$

(e) DeepHit (f) DeepHit $\epsilon = 1$ (g) DeepHit $\epsilon = 4$ (h) DeepHit $\epsilon = 8$

(i) Nnet (j) Nnet $\epsilon = 1$ (k) Nnet $\epsilon = 4$ (l) Nnet $\epsilon = 8$

(m) Coxtime (n) Coxtime $\epsilon = 1$ (o) Coxtime $\epsilon = 4$ (p) Coxtime $\epsilon = 8$

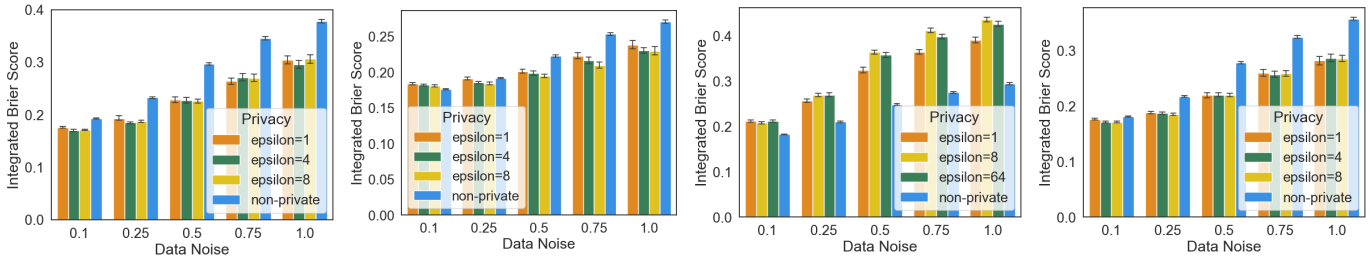
Fig. 8: Predicted Survival Functions for Sample Individuals in METABRIC (best viewed in color): all models were trained over 10 epochs; private models were trained to meet the target ϵ values. Note that survival functions were predicted for the same set of individuals within each row.

and higher C-Index scores. Because private models tend not to overfit/memorize training data, they can be more tolerant of testing samples that come from different distributions. Moreover, private models with weaker privacy guarantees often perform better with noisy testing data. As those models were trained with smaller perturbations, they are likely to achieve a balance between non-overfitting and learning patterns accurately.

It can be seen that when testing data is noisy, non-private discrete-time models (DeepHit and Nnet) have better calibration (i.e., slower degradation in IBS scores) than continuous-time models (DeepSurv and CoxTime). We hypothesize that DeepHit and Nnet may benefit from a smaller domain for time-to-events. In addition, private Nnet models perform poorly in IBS with noisy testing data, as seen in Figure 9c. We note that hyper-parameter tuning for private models may help improve calibration.

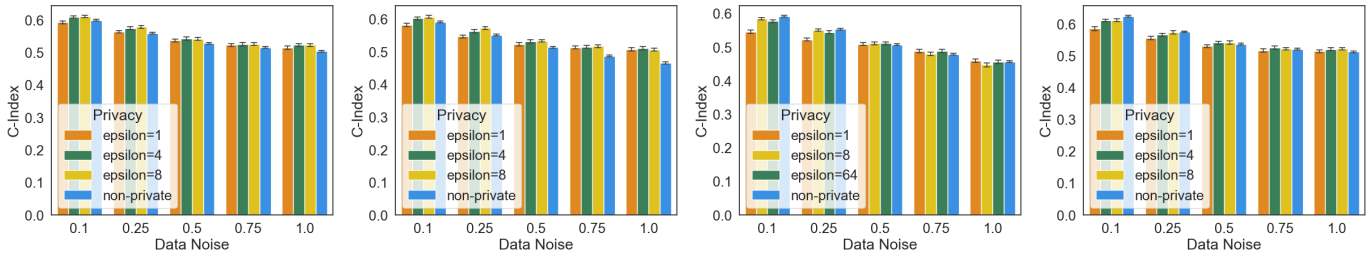
VI. Discussion and Conclusion

We have presented a privacy-protecting approach for sharing useful deep survival models while defending against membership inference. Our approach builds on the rigorous notion of differential privacy, which enables privacy-protecting model sharing without inflicting computation and communication overheads. We have demonstrated that non-private deep survival models leak membership information, and their differentially private counterparts effectively mitigate such



(a) DeepSurv (b) DeepHit (c) Nnet (d) CoxTime

Fig. 9: Integrated Brier Score for Robustness Evaluation with METABRIC (best viewed in color).



(a) DeepSurv (b) DeepHit (c) Nnet (d) CoxTime

Fig. 10: Concordance Index for Robustness Evaluation with METABRIC (best viewed in color).

leakage. We have also shown that differentially private survival models achieve comparable and sometimes better performance (e.g., when testing data is noisy), than that of the non-private models, which may facilitate knowledge sharing in large, diverse collaborative networks.

Our results discovered several opportunities to further improve private deep learning for survival analysis. Firstly, while we demonstrated the leakage of membership inference in deep survival models, the adversarial model in literature can be overly strong, e.g., knowing the training distributions and the target patient's covariates. It would be helpful to investigate a variety of attack models and assumptions to fully understand the range of privacy risks in healthcare research and applications. Future research may consider the adaptation of relaxed privacy notions (e.g., Gaussian differential privacy [19]) to boost the usability of the trained models via improved privacy accountant and amplification [14], [20]. Secondly, differentially private survival models provide effective defense against membership inference, even at weaker levels of theoretical guarantees (e.g., $\epsilon = 8$). That indicates that differential privacy provides stronger privacy protection than practical risks in the context, which may inflict unnecessary utility loss. Future work may bridge the privacy-utility gap by exploring privacy notions and solutions suited for practical risks in the specific domain and application. Lastly, we note that differential privacy may introduce uncertainty in deep learning (as shown by noisy convergence or non-convergence behaviors) and may have different effects on models. Common practices for non-private deep learning, such as parameter tuning, may help mitigate uncertainty effects and deliver high quality models. Existing methods on parameter tuning [21], architecture and feature selection [22], and private model selection [23] may shed light on improving private deep learning for future work.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for valuable feedback. LF has been supported in part by National Science Foundation grants CNS-1951430 and CNS-2144684. LB has been supported in part by National Human Genome Research Institute grants R00HG010493 and RM1HG009034, and a National Library of Medicine grant R01LM013712. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[2] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 945–954, 2018.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 3–18.

[4] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[5] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[6] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[7] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–12, 2018.

[8] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[9] M. F. Gensheimer and B. Narasimhan, "A scalable discrete-time survival model for neural networks," *PeerJ*, vol. 7, p. e6257, 2019.

[10] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression," *Journal of Machine Learning Research*, vol. 20, pp. 1–30, 2019.

[11] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[12] L. Bonomi, X. Jiang, and L. Ohno-Machado, "Protecting patient privacy in survival analyses," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 366–375, 2020.

[13] L. Bonomi and L. Fan, "Sharing time-to-event data with privacy protection," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2022, pp. 11–20.

[14] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled rényi differential privacy and analytical moments accountant," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1226–1235.

[15] N. E. Breslow and N. Chatterjee, "Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 4, pp. 457–468, 1999.

[16] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[17] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statistics in medicine*, vol. 24, no. 24, pp. 3927–3944, 2005.

[18] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.

[19] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 1, pp. 3–37, 2022.

[20] Y. Zhu, J. Dong, and Y.-X. Wang, "Optimal accounting of differential privacy via characteristic function," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 4782–4817.

[21] N. Papernot and T. Steinke, "Hyperparameter tuning with renyi differential privacy," in *International Conference on Learning Representations*, 2021.

[22] W. Bao, L. A. Bauer, and V. Bindschaedler, "On the importance of architecture and feature selection in differentially private machine learning," *arXiv preprint arXiv:2205.06720*, 2022.

[23] L. Fan and A. Pokkunuru, "Dpnet: Differentially private network traffic synthesis with generative adversarial networks," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2021, pp. 3–21.