Challenges of Differentially Private Prediction in Healthcare Settings

Vinith M. Suriyakumar[1*], **Nicolas Papernot**[1], **Anna Goldenberg**[1] and **Marzyeh Ghassemi**[1]
[1]University of Toronto, Vector Institute
vinith@cs.toronto.edu

Abstract

Privacy-preserving machine learning is becoming increasingly important as models are being used on sensitive data such as electronic health records. Differential privacy is considered the gold standard framework for achieving strong privacy guarantees in machine learning. Yet, the performance implications of learning with differential privacy have not been characterized in the presence of time-varying hospital policies, care practices, and known class imbalance present in health data. First, we demonstrate that due to the long-tailed nature of healthcare data, learning with differential privacy results in poor utility tradeoffs. Second, we demonstrate through an application of influence functions that learning with differential privacy leads to disproportionate influence from the majority group on model predictions which results in negative consequences for utility and fairness. Our results highlight important implications of differentially private learning; which focuses by design on learning the body of a distribution to protect privacy but omits important information contained in the tails of healthcare data distributions.

1 Introduction

The potential for machine learning to improve healthcare through secondary analysis of electronic health records (EHR) data has been demonstrated across a variety of tasks [Tomašev *et al.*, 2019; Gulshan *et al.*, 2016; Wu *et al.*, 2019; Rajkomar *et al.*, 2018b]. To protect patient information, EHR data is often anonymized. However, linkage attacks allow a malicious entity to leverage access to external data to de-anonymize data [Narayanan and Shmatikov, 2008]. This external data may for instance contain auxiliary information about individuals whose data was anonymized in the first place. Linkage attacks have been used to de-anonymize public releases of healthcare data [Sweeney, 2015]. Machine learning models are also known to be susceptible to membership inference and attribute inference attacks [Shokri *et al.*, 2017] where the adversary may recover sensitive information about an individual contained in the training set of a model trained without privacy guarantees.

Differential privacy (DP) is the current gold standard framework for analyzing the privacy guarantees of randomized algorithms, of which machine learning algorithms are an example of [Dwork *et al.*, 2006]. Differentially private stochastic gradient descent (DP-SGD) is a commonly-used technique for training machine learning models with differential privacy [Abadi *et al.*, 2016]. Despite its advantages in terms of privacy guarantees provided for the training data, the DP-SGD training algorithm often introduces a privacy-utility tradeoff. This tradeoff has been characterized in settings such as vision [Papernot *et al.*, 2020] and tabular data [Jayaraman and Evans, 2019]. The utility tradeoff is highly dataset dependent, motivating our novel, extensive analysis of its use in a healthcare setting.

EHR data is different from other domains because it often contains noisy and missing measurements, changing populations, and changing healthcare practices that result in dataset shift over time. Furthermore, rare positive cases and minorities often exist in the "tails". These issues present unique challenges ensuring strong performance in the presence of class imbalance for existing machine learning methods. However, DP-SGD has been primarily developed and evaluated on datasets that are well-balanced. Because DP-SGD focuses on learning the body of a distribution as the privacy increases, the amount of information lost from the tails of healthcare data distributions is important to understand.

Given these unique challenges in the data, we examine whether DP-SGD can feasibly be used in machine learning for healthcare. We train both linear models and neural networks on several tasks in MIMIC-III, a publicly available EHR database. We analyze **(1)** whether linear models or neural networks provide better tradeoffs and **(2)** whether influence functions can be used to explain the tradeoffs and whether DP results in model decision making that is concerning based on known clinical practices.

We demonstrate that DP-SGD is currently not suited for use in machine learning for healthcare because it focuses on the body of the data distribution and (by design, to protect their privacy) does not learn about the tails of the data distribution. For instance, we find that learning with strong differential privacy guarantees can cause the training data for a majority ethnicity group to have more influence on the model's

[*] Contact Author

predictions. This is an undesirable effect given that the use of ethnicity remains poorly understood in healthcare. Future work should target relaxations of differential privacy that enable improved learning on long-tailed distributions and modifications to DP-SGD that address its feasibility for machine learning in healthcare.

2 Related Work

Differential Privacy. Differential privacy formalizes the privacy guarantees of randomized algorithms, such as stochastic gradient descent. [Dwork *et al.*, 2006]. An algorithm is differentially private if its output is statistically indistinguishable when applied to two input datasets that differ by only one record (Hamming distance of 1). Formally, a learning algorithm L that trains models from the dataset D satisfies (ϵ,δ)-DP if the following holds for all training datasets d and d' with a Hamming distance of 1:

$$Pr[L(d) \in D] \leq e^{\epsilon} Pr[L(d') \in D] + \delta \qquad (1)$$

The parameter ϵ measures the formal privacy guarantee by defining a strong upper bound on the privacy loss in the worst possible case. A smaller ϵ represents stronger privacy guarantees. The δ factor allows for some probability that the property may not hold.

Differential Privacy in Healthcare. Prior work on DP in machine learning for health has focused on the distributed setting, where multiple hospitals collaborate to learn a model [Beaulieu-Jones *et al.*, 2018; **?**], and has primarily found that DP learning negatively impacts model AUROC. We instead focus on analyzing the tradeoffs in multiple healthcare tasks of varying levels of class imbalance and providing an empirical explanation for these tradeoffs using influence functions with an emphasis on the impact that DP has on subgroups.

Explaining Machine Learning Through Influence. Initial findings show that an important factor in clinicians trusting machine learning is presenting them an explanation of how the model came to its prediction [Tonekaboni *et al.*, 2019]. Influence functions have been demonstrated as a technique to measure the effects of individual training points on a model's prediction [Koh and Liang, 2017] and have been extended to approximate the effects of subgroups on a model's prediction [Koh *et al.*, 2019]. Recent work demonstrates that memorization is required for small generalization error on long-tailed distributions and that influence function can be used to explain this phenomena in learning algorithms [Feldman, 2020]. We use these findings to inform why healthcare has poor utility tradeoffs and whether DP-SGD makes predictions in ways that clinicians would trust.

3 Data

For our healthcare tasks, we use the MIMIC-III database [Johnson *et al.*, 2016]—a publicly available anonymized EHR dataset of intensive care unit (ICU) patients.

4 Methods

We define two distributions of patients, p and q where q is shifted and assume the ability to sample from both. Given a dataset $\{(x_1, y_1), ...(x_n, y_n)\} \sim p$, test data from our shifted distribution $\{(x_1, y_1), ...(x_n, y_n)\} \sim q$, and three levels of DP $\{None, Low, High\}$, we analyze the tradeoffs between privacy and utility. In the healthcare setting, x_i is a combination of static variables and time series for each patient. We partition the healthcare data into p and q based on the year of care: for a given year y, all records prior to y are used to train, and records from y itself are test. We use linear models and neural networks for experiments on two binary prediction tasks and one mutliclass prediction task.

Models For all healthcare tasks analyses, we choose one linear model and one neural network per task, based on the best baselines outlined in prior work creating benchmarks for the MIMIC-III dataset [Wang *et al.*, 2019]. For binary prediction tasks we use logistic regression (LR) [Cox, 1972] and gated recurrent unit with decay (GRUD) [Che *et al.*, 2018]. For our multiclass prediction task, we use LR and 1D CNNs.

Differentially Private Training We train models without privacy guarantees using stochastic gradient descent (SGD). When training models with privacy guarantees, we use DP-SGD [Abadi *et al.*, 2016]. The modifications made to SGD involve clipping gradients computed on a per-example basis to have a maximum ℓ_2 norm, and then adding Gaussian noise to these gradients before applying the model parameter updates [Abadi *et al.*, 2016]. We choose three different levels of privacy to measure the effect of increasing levels of privacy outlined below:

Privacy Level	None	Low	High
(Clip Norm, Noise Multiplier)	(0.0,0.0)	(5.0,0.1)	(1.0,1.0)

Table 1: Clipping norm and noise multiplier values used to achieve our high and low privacy settings.

Influence Functions Given a set of training points $\{x_i, y_i\} \sim p$ and model parameters θ we calculate the influence through the loss to be $L(x, \theta)$ and let $\frac{1}{n}\Sigma_{i=1}^{n} L(x_i, \theta)$. The empirical risk is assumed to be smooth and strictly convex in θ. Given this setup and assumptions we focus our analysis on logistic regression since our neural networks are non-convex in θ. Using the approach from [Koh and Liang, 2017] we analyze the influence of all training points on the loss for each test point defined in Equation 2. Influence functions use an additive property for interpreting the influence of subgroups showing that the group influence is the sum of the influences of all individual points in the subgroup but that this is usually an underestimate of the true influence of removing the subgroup [Koh *et al.*, 2019].

$$I_{up,loss}(z_{train}, z_{test}) = -\nabla_{\theta} L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z_{train}, \hat{\theta}) \qquad (2)$$

Experimental Design First, we analyze the privacy-utility tradeoff by training linear models and neural networks with three privacy levels in the healthcare tasks. We measure the utility using AUC in the healthcare tasks, with the privacy measured using the ϵ bound derived from DP. Finally, we use influence functions to measure the influence of points from p on the test loss of points from q and the connection between these influences and the utility tradeoff. Influences above 0 mean the training point was helpful in minimizing the test point loss and below 0 mean the training point increased the test point loss.

5 Healthcare Tasks Have Worse Utility Tradeoffs

We characterize the privacy-utility tradeoff across healthcare tasks varying tail sizes. For the healthcare tasks, we define the privacy utility tradeoffs by taking the average AUROC across all years of care, for each of the privacy levels and models defined in the methods in Table 2.

Task	Model	None	Low	High
Mortality	LR	$0.82 \pm 0.03\ (\infty)$	$0.76 \pm 0.05\ (3.50 \cdot 10^5)$	$0.60 \pm 0.04\ (3.54)$
	GRUD	$0.79 \pm 0.03\ (\infty)$	$0.59 \pm 0.09\ (1.59 \cdot 10^5)$	$0.53 \pm 0.03\ (2.65)$
Length of Stay	LR	$0.69 \pm 0.02\ (\infty)$	$0.66 \pm 0.03\ (3.50 \cdot 10^5)$	$0.60 \pm 0.04\ (3.54)$
	GRUD	$0.67 \pm 0.03\ (\infty)$	$0.63 \pm 0.02\ (1.59 \cdot 10^5)$	$0.61 \pm 0.03\ (2.65)$
Intervention Onset (Vaso)	LR	$0.90 \pm 0.03\ (\infty)$	$0.87 \pm 0.03\ (1.63 \cdot 10^7)$	$0.77 \pm 0.05\ (0.94)$
	CNN	$0.88 \pm 0.04\ (\infty)$	$0.86 \pm 0.02\ (5.95 \cdot 10^7)$	$0.68 \pm 0.04\ (0.66)$

Table 2: Privacy utility tradeoff across healthcare tasks. The healthcare tasks have a significant tradeoff between the High and Low or None setting. The tradeoff is better in more balanced tasks (length of stay and intervention onset), and worst in tasks such as mortality where class imbalance is present at 7.4% positive cases. There is a 22% and 26% drop in the AUROC between no privacy and high privacy settings for mortality prediction for LR and GRUD respectively. We provide the ϵ guarantees in parentheses.

Comparing the privacy utility tradeoffs in Table 2, DP-SGD has negative consequences on the model utility when the task and dataset is difficult due to important tails in the data distribution. The extreme tradeoffs in mortality prediction capture this issue since the positive cases are in the tails of the distribution.

6 Group Privacy Gives Over-Influence for Majority Groups

We focus group privacy analyses on the LR model for mortality prediction, examining the no privacy and high privacy settings. Along with guaranteeing individual privacy, DP guarantees group privacy. The group privacy guarantees state that the ϵ guarantee degrades linearly based on the size of the group. Gradient clipping results in tightly bounded influence of all training points across test points whereas individual training points have much more helpful and harmful influence without DP.

Utility Tradeoff The mortality task, where tails of the distribution hold the patients who died, has the largest privacy-utility tradeoff because DP focuses on the patients who survived that make up the body of the distribution. This results in over-influence for the patients who survived, while the no privacy model finds the patients who died to be the most helpful in its predictions (Fig. 1 and Table 3).

This result requires careful thought, as the tails of the label distribution are minority-rich, which results in poor performance for DP. Differences in access, practice, or recording reflect societal biases [Rajkomar *et al.*, 2018a; Rose, 2018], and models trained on biased data may exhibit unfair performance in populations due to this underlying variation [Chen *et al.*, 2019]. Further, while patients with the same diagnosis are usually more helpful for estimating prognosis in practice [Croft *et al.*, 2015], labels in healthcare often lack precision or, in some cases, may be unreliable [O'malley *et al.*, 2005]. In this setting, understanding what factors are consistent in patient phenotypes is an important task [Halpern *et al.*, 2016; Yu *et al.*, 2017].

Fairness Tradeoff We use influence functions to approximate the group influences of different ethnicities to understand the privacy fairness tradeoffs. Following the same setting as the previous section, we present the group influences of different ethnicities in the training set on the test loss in Fig. 2.

Group privacy results in white patients having a more significant influence, both helpful and harmful, on both white and black patients in the high privacy setting (Table 4). Ethnicity is currently an important consideration in clinical practice, where different risk profiles are often assumed for the patients of different races [Martin, 2011]. The validity of this stratification has recently been called into question by the clinical community, and is still being explored [Eneanya *et al.*, 2019]. Prior work has established the complexity of treatment variation in practice, as patient care plans are highly individualized, e.g., in a cohort of 250 million patient, 10% of diabetes and depression patients and almost 25% of hypertension patients had a unique treatment pathway [Hripcsak *et al.*, 2016]. Thus having the white patients be most influential in the predictions of the black patients in models trained with different privacy constraints, should be carefully considered.

7 Conclusion

In this work, we investigate the feasibility of using DP-SGD to train models for healthcare prediction tasks. We find that DP-SGD is not well-suited to healthcare prediction tasks in its current formulation. First, we demonstrate that DP-SGD increasingly targets the body of a distribution as privacy level increases, losing important information about minority classes (e.g., dying patients, minority ethnicities) that lies in the distributional tails. Our analyses demonstrate that this results in extreme privacy-utility tradeoffs. We show that the group privacy guarantee of DP plays a large role in this tradeoff, and that it results in over-influence of the majority group (e.g., white patients, healthy patients) on patients in the minority class label (hospital mortality) and ethnicity (e.g., black patients). This imposed asymmetric valuation of data by the model requires careful thought, because the appropriateness of minority class membership use in clinical settings in an active topic of discussion and debate. Future work should target modifying DP-SGD, or creating novel DP learning algorithms, that can learn from data distribution tails effectively, without compromising privacy.

Privacy Level	Average Majority Influence	Average Minority Influence	Most Helpful Group	Most Harmful Group Influence
None	-1.07 ± 7.25	2.28 ± 6.91	Died (Minority)	Survived (Majority)
Low	-0.34 ± 0.95	0.03 ± 0.18	Survived (Majority)	Survived (Majority)
High	-0.14 ± 4.69	0.04 ± 1.34	Survived (Majority)	Survived (Majority)

Table 3: Group influence summary statistics of training data by class label in all privacy levels for all test patients. Privacy changes the most helpful group from the minority to the majority and minimizes the minority group's helpful influence.

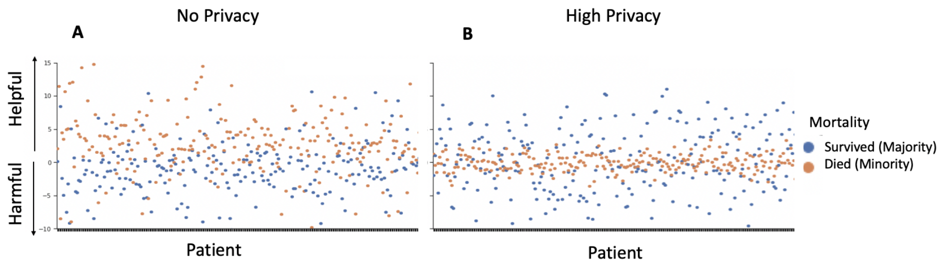

Figure 1: Group influence of training data per class label in no privacy (A) and high privacy (B) settings on 100 test patients with highest influence variance. In the no privacy setting, patients who died have a helpful influence despite being a minority class. High privacy gives the majority group the most influence due to the group privacy guarantee. This results in prognostically dissimilar patients having the most influence on the model's prediction.

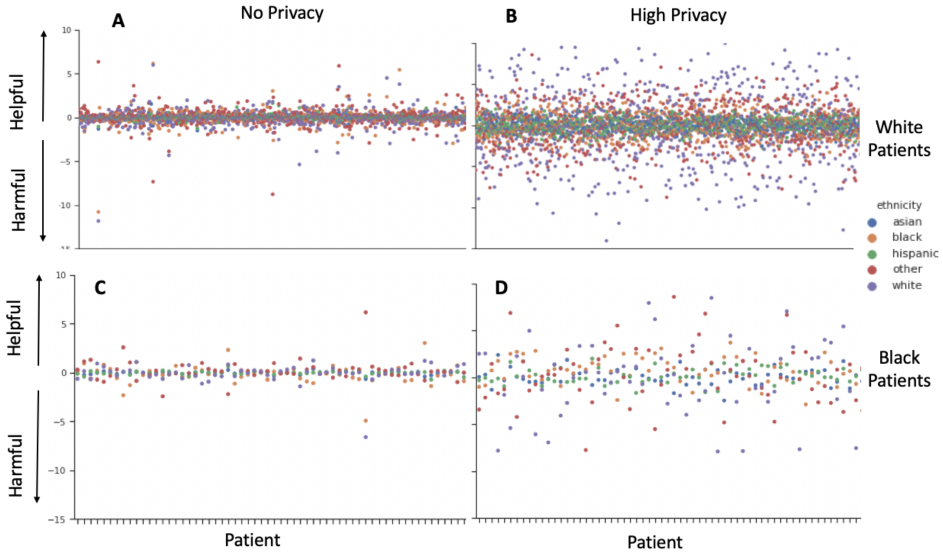

Figure 2: Group influence of training data per ethnic groups on 100 test patients with highest influence variance. The group influence of our majority ethnicity (white patients) is enhanced significantly in the high privacy setting, as demonstrated by the increased amplitude of those points in (B) and (D). In the no privacy setting the group influence of each ethnicity is similar for both white (A) and black patients (C).

White Test Patients				
Privacy Level	Average White Influence	Average Black Influence	Most Helpful Ethnicity	Most Harmful Ethnicity
None	0.29 ± 2.40	0.71 ± 1.40	White (Majority)	White (Majority)
Low	-0.22 ± 0.70	-0.03 ± 0.17	White (Majority)	White (Majority)
High	-0.11 ± 3.94	0.03 ± 1.35	White (Majority)	White (Majority)
Black Test Patients				
Privacy Level	Average White Influence	Average Black Influence	Most Helpful Ethnicity	Most Harmful Ethnicity
None	0.48 ± 1.39	0.44 ± 2.19	Black (Minority)	White (Majority)
Low	-0.23 ± 0.75	-0.03 ± 0.18	White (Majority)	White (Majority)
High	-0.40 ± 4.10	0.12 ± 1.45	White (Majority)	White (Majority)

Table 4: Group influence summary statistics of white and black training patients for all privacy levels for all white and black test patients. for the black test patients, privacy changes the most helpful group from black patients to the majority white patients and minimizes black patients' helpful influence. This needs careful consideration as the use of ethnicity is still being investigated in clinical practice.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, 2016. ACM.

[Beaulieu-Jones *et al.*, 2018] Brett K Beaulieu-Jones, William Yuan, Samuel G Finlayson, and Zhiwei Steven Wu. Privacy-preserving distributed deep learning for clinical data. *arXiv preprint arXiv:1812.01484*, 2018.

[Che *et al.*, 2018] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):6085, April 2018.

[Chen *et al.*, 2019] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 21(2):167–179, 2019.

[Cox, 1972] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[Croft *et al.*, 2015] Peter Croft, Douglas G Altman, Jonathan J Deeks, Kate M Dunn, Alastair D Hay, Harry Hemingway, Linda LeResche, George Peat, Pablo Perel, Steffen E Petersen, et al. The science of clinical practice: disease diagnosis or patient prognosis? evidence about "what is likely to happen" should shape clinical practice. *BMC medicine*, 13(1):20, 2015.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[Eneanya *et al.*, 2019] Nwamaka Denise Eneanya, Wei Yang, and Peter Philip Reese. Reconsidering the consequences of using race to estimate kidney function. *Jama*, 322(2):113–114, 2019.

[Feldman, 2020] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[Gulshan *et al.*, 2016] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, December 2016.

[Halpern *et al.*, 2016] Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–740, 2016.

[Hripcsak *et al.*, 2016] G. Hripcsak, P.B. Ryan, J.D. Duke, N.H. Shah, R.W. Park, V. Huser, M.A. Suchard, M.J. Schuemie, F.J. DeFalco, A. Perotte, et al. Characterizing treatment pathways at scale using the ohdsi network. *Proceedings of the National Academy of Sciences*, 113(27):7329–7336, 2016.

[Jayaraman and Evans, 2019] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1895–1912, 2019.

[Johnson *et al.*, 2016] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, May 2016.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

[Koh *et al.*, 2019] Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems*, pages 5255–5265, 2019.

[Martin, 2011] Toni Martin. The color of kidneys. *American Journal of Kidney Diseases*, 58(5):A27–A28, 2011.

[Narayanan and Shmatikov, 2008] A Narayanan and V Shmatikov. Robust de-anonymization of large sparse datasets [netflix]. In *IEEE Symposium on Research in Security and Privacy, Oakland, CA*, 2008.

[O'malley *et al.*, 2005] K.J. O'malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, and C.M. Ashton. Measuring diagnoses: ICD code accuracy. *Health Services Research*, 40(5p2):1620–1639, 2005.

[Papernot *et al.*, 2020] Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy, 2020.

[Rajkomar *et al.*, 2018a] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[Rajkomar *et al.*, 2018b] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

[Rose, 2018] Sherri Rose. Machine learning for prediction in electronic health data. *JAMA network open*, 1(4):e181404–e181404, 2018.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017*

IEEE Symposium on Security and Privacy (SP), pages 3–18. IEEE, 2017.

[Sweeney, 2015] Latanya Sweeney. Only you, your doctor, and many others may know. *Technology Science*, 2015092903(9):29, 2015.

[Tomašev *et al.*, 2019] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, August 2019.

[Tonekaboni *et al.*, 2019] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134*, 2019.

[Wang *et al.*, 2019] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Michael C. Hughes, Tristan Naumann, and Marzyeh Ghassemi. MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III. *arXiv:1907.08322 [cs, stat]*, July 2019. arXiv: 1907.08322.

[Wu *et al.*, 2019] Denny Wu, Hirofumi Kobayashi, Charles Ding, Lei Cheng, and Keisuke Goda Marzyeh Ghassemi. Modeling the Biological Pathology Continuum with HSIC-regularized Wasserstein Auto-encoders. *arXiv:1901.06618 [cs, stat]*, January 2019. arXiv: 1901.06618.

[Yu *et al.*, 2017] S. Yu, Y. Ma, J. Gronsbell, T. Cai, A.N. Ananthakrishnan, V.S. Gainer, S.E. Churchill, P. Szolovits, S.N. Murphy, I.S. Kohane, et al. Enabling phenotypic big data with phenorm. *Journal of the American Medical Informatics Association*, 25(1):54–60, 2017.