Proceedings of Machine Learning Research 133:206–215, 2021 NeurIPS 2020 Competition and Demonstration Track

Hide-and-Seek Privacy Challenge: Synthetic Data Generation *vs.* Patient Re-identification

James Jordon JAMES.JORDON@WOLFSON.OX.AC.UK
University of Oxford

Daniel Jarrett DANIEL.JARRETT@MATHS.CAM.AC.UK
University of Cambridge

Evgeny Saveliev ES583@CAM.AC.UK
University of Cambridge

Jinsung Yoon JINSUNGYOON@GOOGLE.COM
Google Cloud AI
University of California, Los Angeles

Paul Elbers P.ELBERS@AMSTERDAMUMC.NL
Amsterdam UMC

Patrick Thoral P.THORAL@AMSTERDAMUMC.NL
Amsterdam UMC

Ari Ercole AE105@CAM.AC.UK
University of Cambridge
Cambridge University Hospitals NHS Foundation Trust

Cheng Zhang CHENG.ZHANG@MICROSOFT.COM
Microsoft Research

Danielle Belgrave DANIELLE.BELGRAVE@MICROSOFT.COM
Microsoft Research

Mihaela van der Schaar MV472@CAM.AC.UK
University of Cambridge
University of California, Los Angeles
Alan Turing Institute

Editors: Hugo Jair Escalante and Katja Hofmann

Abstract

The *clinical time-series* setting poses a unique combination of challenges to data modelling and sharing. Due to the high dimensionality of clinical time series, adequate de-identification to preserve privacy while retaining data utility is difficult to achieve using common de-identification techniques. An innovative approach to this problem is *synthetic data generation*. From a technical perspective, a good generative model for time-series data should preserve temporal dynamics; new sequences should respect the original relationships between high-dimensional variables across time. From the privacy perspective, the model should prevent *patient re-identification*. The NeurIPS 2020 Hide-and-Seek Privacy Challenge was a novel two-tracked competition to simultaneously accelerate progress in tackling both problems. In our head-to-head format, participants in the generation track ("hiders") and the patient re-identification track ("seekers") were directly pitted against each other by way of a new, high-quality intensive care time-series dataset: the Amsterda-

mUMCdb dataset. In this paper we present an overview of the competition design, as well as highlighting areas we feel should be changed for future iterations of this competition.
Keywords: Clinical Time-series Data; Data Privacy; Synthetic Data Generation; Patient Re-identification; Membership Inference Attack.

1. Introduction

Coupled with advances in machine learning, the vast quantities of clinical data now stored in machine-readable form have the potential to revolutionize healthcare. At the same time, this enterprise is threatened by the fact that patient data are inherently highly sensitive, and privacy concerns have recently been thrown into sharp relief by several high-profile data breaches that have greatly undermined public confidence (see e.g. Shah (2017); Price and Cohen (2019)). We seek novel methods capable of bridging the gap between data-hungry techniques in machine learning and privacy-conscious applications in healthcare settings.

The Vision for Synthetic Data Perhaps the most attractive use cases for synthetic data are that of *synthetic data clearing houses* and for running *algorithm-finding competitions* (in the same spirit as this competition was run). Synthetic data clearing houses are an attractive idea that involve setting up an institution responsible for the generation of synthetic data on the basis of real data. Such institutions would then be entrusted by existing data holders to generate synthetic data for a variety of use-cases that could then be made publicly available, allowing the ML community at large to push advancements for a wide range of problems. Closely related is the idea that synthetic data could be used for the purpose of finding (the) group(s) that are most appropriate for developing a model for a given task (Jordon et al., 2018). Such synthetic data would need to be reflective of the real data in terms of model performance for the task at hand, but can be generated at varying levels of privacy as participants advance through the competition, with the winner(s) being given access to the full data to develop the final models.

Clinical Time-series Central among this balancing act is the development of techniques for modeling and sharing synthetic patient records in lieu of real data. However, the setting of clinical time-series data poses a unique combination of challenges to data modeling and sharing. From a technical perspective, the learning problem in question is one of *synthetic data generation*—a good generative model for time-series data should preserve temporal dynamics, in the sense that new sequences respect the original relationships between high-dimensional variables across time. Simultaneously from a social perspective, the privacy problem in question is one of *patient re-identification*—a good generation technique should have the effect of preserving membership privacy, in the sense that the algorithm limits vulnerability of individual training instances to the risk of membership inference attacks.

Time-series Generation Purely from the standpoint of generative modeling, the *sequential* setting of clinical time-series data already presents a distinctive learning challenge. A model is not only tasked with capturing the distributions of patient features at each point in time, but it should also adequately reflect the potentially complex evolution of those variables over time. Existing methods directly apply the generative adversarial network (GAN) framework to temporal data, primarily by instantiating recurrent neural network (RNN)

models as generators and discriminators (e.g. Esteban et al. (2017); Mogren (2016); Ramponi et al. (2018)). Such straightforward approaches neglect to leverage the autoregressive prior, and have been shown insufficient for ensuring that the network dynamics efficiently capture stepwise dependencies in the original training data (Yoon et al. (2019)).

Privacy and Identification Most importantly, the question of synthetic data generation cannot be divorced from concerns of *privacy*. While de-identified data are commonly used for model development, existing notions of anonymity are often limited in scope: k-anonymity, l-diversity, and τ-closeness are only aimed at protecting "sensitive" data (e.g. diagnoses) from an attack on a small number of quasi-identifiers (see Sweeney (2002); Machanavajjhala et al. (2007) and Li et al. (2007) respectively), and differential privacy (see Dwork et al. (2014)) does not directly correspond to well-understood notions of leakage—such as vulnerability to membership inference attacks. In other directions, attempts to match GDPR with a mathematical notion of privacy have been attempted (Yoon et al., 2020), though whether the notion properly aligns with GDPR or whether the defined notion is robust to privacy attacks has not been investigated thoroughly. In practice, the risk of patient re-identification is a pressing concern: Consider a rogue insurance company discriminating against high-risk patients per financial incentive. Such concerns caution medical institutions against releasing data for public research, hampering progress in the validation of novel computational models for real-world clinical applications.

Hide-and-Seek Challenge The NeurIPS 2020 Hide-and-Seek Privacy Challenge[1] was a novel *two-tracked* competition to simultaneously accelerate progress in tackling both problems. In our head-to-head format, participants in the synthetic data generation track ("hiders") and the patient re-identification track ("seekers") were directly pitted against each other: The latter submitted methods for launching membership inference attacks, while the former submitted methods for synthesizing patient data that are robust to such attacks—all while maintaining faithfulness to the original data. Importantly, rather than falling back on fixed theoretical notions of anonymity, we allowed participants on both sides to uncover the best approaches *in practice* for launching or defending against privacy attacks. We sought to advance generative techniques for dense, high-dimensional temporal data streams that are clinically meaningful in terms of fidelity and predictivity, as well as capable of minimizing privacy risks in terms of the concrete notion of patient re-identification.

2. Two-Track Format

The competition involved a *two-sided platform* (hosted on CodaLab) for synthetic data generation and patient re-identification methods to compete among and against each other. Participants were invited to compete in either or both tracks of the interactive challenge: (1) the *hider* (synthetic data generation) track, and (2) the *seeker* (re-identification) track.

2.1. Hider track

In the generation track, participants were tasked with developing an algorithm that generates synthetic data. Submissions were an algorithm $\mathcal{M}_G$ (i.e. not just a trained model),

1. https://www.vanderschaar-lab.com/privacy-challenge/

whose **input** will be random subsets of an unseen subset of the dataset,

$$\mathcal{D}_{real} \subset \mathcal{D}\,,$$

and whose **output** is a synthetic dataset,

$$\mathcal{D}_{syn} = \mathcal{M}_G(\mathcal{D}_{real})\,.$$

2.2. Seeker Track

In the *patient re-identification* track, participants were tasked with developing an algorithm that performs membership inference on synthetic data generation algorithms. Submissions were an algorithm, $\mathcal{M}_R$ (which may contain trained models from the public data), whose **input** should be tuples of the form

$$(\mathcal{M}_G, \mathcal{D}_{syn}, \mathcal{D}^{enl}_{real}) \tag{1}$$

where $\mathcal{M}_G$ is an indicator for the generation algorithm used, $\mathcal{D}_{syn}$ is generated by $\mathcal{M}_G$ and $\mathcal{D}^{enl}_{real}$ is a random subset of $\mathcal{D}$ that contains the real data, $\mathcal{D}_{real}$, used to generate $\mathcal{D}_{syn} = \mathcal{M}_G(\mathcal{D}_{real})$ (i.e. a randomly *enlarged* copy of $\mathcal{D}_{real}$). The **output** must be a classification of each element of $\mathcal{D}^{enl}_{real}$, in which the goal is to classify the elements as:

'*in* $\mathcal{D}_{real}$' or '*not in* $\mathcal{D}_{real}$'

or equivalently the output must be a subset $\mathcal{D}_{pred} \subset \mathcal{D}^{enl}_{real}$ corresponding to the elements the algorithm classifies as being '*in* $\mathcal{D}_{real}$'.

2.3. Scoring and Ranking

Suppose $\mathcal{M}^1_G, ..., \mathcal{M}^{N_G}_G$ are submissions to the generation track and $\mathcal{M}^1_R, ..., \mathcal{M}^{N_R}_R$ are submissions to the re-identification track. Let $\mathcal{D}^i_{real} \subset \mathcal{D}^{enl\ i}_{real} \subset \mathcal{D}_{priv}$ be 10 random subsets (and enlarged subsets) of the private data. For each generation algorithm and each real dataset we generate a synthetic dataset $\mathcal{D}^{i,j}_{syn} = \mathcal{M}^j_G(\mathcal{D}^i_{real})$.

2.3.1. Evaluating Hiders

Evaluating hiders is not a straightforward task. There is a natural trade-off between privacy and fidelity - any submission will demonstrate a certain level of privacy and a certain level of fidelity, and as one increases the other likely decreases. In order to determine a ranking for hider algorithms we were required to translate these two quantities into a single comparable quantity. In this instance, we chose to fix a minimum fidelity threshold that hider algorithms must exceed in order to qualify for ranking. Ranking was then performed according to the privacy attained by each submission. This makes the question we asked *"How private can we make the data while maintaining usefulness?"* rather than *"How useful can we make the data while maintaining privacy?"*. Both questions have merit, and in an ideal world the trade-off would be explored more thoroughly, but as a first competition in this space, we chose to keep the metric simple so as to ensure the competition was as accessible as possible.

Utility threshold To measure the quality of a generated synthetic dataset, $\mathcal{D}_{syn}$, we use the train-on-synthetic-test-on-real paradigm (Esteban et al., 2017). Abstractly, we consider a task, $\mathcal{T}$, that can be performed on the dataset $\mathcal{D}$, which can be performed by some algorithm $\mathcal{M}_{\mathcal{T}}$ and for which there is an increasing performance metric $\mathcal{C}_{\mathcal{T}}$ that maps an algorithm, train data, test data tuple to a scalar performance value, $(\mathcal{M}_{\mathcal{T}}, \mathcal{D}_{train}, \mathcal{D}_{test}) \mapsto C \in \mathbb{R}$. A synthetic dataset's utility score for task $\mathcal{T}$ is then defined by

$$\mathcal{S}_{\mathcal{T}}(\mathcal{D}_{syn}) = \frac{\mathcal{C}_{\mathcal{T}}(\mathcal{M}_{\mathcal{T}}, \mathcal{D}_{syn}, \mathcal{D}^{test}_{real})}{\mathcal{C}_{\mathcal{T}}(\mathcal{M}_{\mathcal{T}}, \mathcal{D}^{train}_{real}, \mathcal{D}^{test}_{real})} \tag{2}$$

which we can require is above some minimum threshold. Given several tasks, $\mathcal{T}_1, ..., \mathcal{T}_{n_{\mathcal{T}}}$, we required that a dataset passes all tasks. For the competition we considered 2 tasks that the synthetic data must be suitable for: (1) feature prediction, and (2) sequential prediction.

Feature prediction For the first task we randomly selected 10 features. For each of these features, the task was to predict this feature given all of the remaining other features, which we denote by $\mathcal{T}_l, l = 1, ..., 10$. As metric, $\mathcal{C}_{\mathcal{T}}$, we use $\frac{1}{\text{RMSE}}$ for continuous features (to create an increasing performance metric), AUROC for binary and accuracy for categorical.

Sequential prediction For the second task, the goal is to perform 1-step-ahead prediction, denoted by $\mathcal{T}_{11}$. The prediction model will predict the feature values at time step $t+1$ based on the entire history up to and including time step t. As metric we use the sum of errors across the different features, using $\frac{1}{\text{RMSE}}$ for continuous features (again to create an increasing performance metric), AUROC for binary and accuracy for categorical.

Passing the threshold In order for an algorithm, $\mathcal{M}^j_G$, to pass the threshold, the following must hold

$$\mathcal{S}_{\mathcal{T}_l}(\mathcal{D}^{i,j}_{syn}) > p \text{ for all } j \in \{1, ..., 10\}, l \in \{1, ..., 11\} \tag{3}$$

for some threshold value p. In the competition, we set $p = 0.2$, where all metrics involved were RMSE and thus this corresponded to requiring that the RMSE produced by synthetic data was at most 5 times larger than the RMSE produced by the real data.

Privacy Ranking Models that pass the utility threshold were ranked according to their robustness to re-identification (the quality scores are no longer relevant once the bar has been passed). The re-identification score, $\mathcal{R}^j$, of generation algorithm $\mathcal{M}^j_G$ is given by

$$\mathcal{R}^j = \max_k \sum_{i=1}^{10} \mathcal{S}^{i,j,k} \tag{4}$$

where $\mathcal{S}^{i,j,k}$ is defined below. We note that a *lower* score is better. This score corresponds to how well the *best* performing re-identification algorithm (for the given generation model) is able to re-identify *on average* across the different synthetic datasets generated by $\mathcal{M}_G$.

2.3.2. Evaluating Seekers

For each $i = 1, ..., 10$ and $j = 1, ..., N_G$, a re-identification algorithm, $\mathcal{M}_R^k$, is assigned a score $\mathcal{S}^{i,j,k} \in [0, 1]$ according to its classification accuracy, given by:

$$\mathcal{S}^{i,j,k} = \frac{|\mathcal{D}_{pred}^{i,j,k} \cap \mathcal{D}_{real}^{i}| + |\mathcal{D}_{pred}^{i,j,k\,c} \cap \mathcal{D}_{real}^{i\,c}|}{|\mathcal{D}_{real}^{enl\,i}|} \tag{5}$$

where ^{c} denotes the compliment of the set (within $\mathcal{D}_{real}^{enl\,i}$) and $\mathcal{D}_{pred}^{i,j,k} = \mathcal{M}_R^k(\mathcal{M}_G^j, \mathcal{D}_{syn}^{i,j}, \mathcal{D}_{real}^{enl\,i})$.

To create an overall score for a re-identification algorithm, we will average the score of the algorithm across the 10 synthetic datasets for *each* generation algorithm *that passed the utility threshold*. Let $\mathcal{P}_G = \{j : \mathcal{S}_{\mathcal{T}_l}(\mathcal{D}_{syn}^{i,j}) > p, \forall j = 1, ..., 10, \forall l = 1, ..., 11\}$ be the indexing set of hiders that passed the utility bar. The overall score, $\mathcal{S}_O$, of re-identification algorithm, $\mathcal{M}_R^k$, is given by

$$\mathcal{S}_O^k = \frac{1}{10N_G} \sum_{i \in \mathcal{P}_G} \sum_{j=1}^{10} \mathcal{S}^{i,j,k}\,, \tag{6}$$

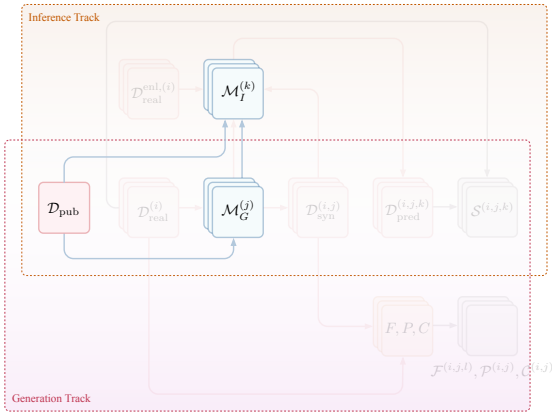
where for the seekers, a *higher* score is better.

3. Dataset

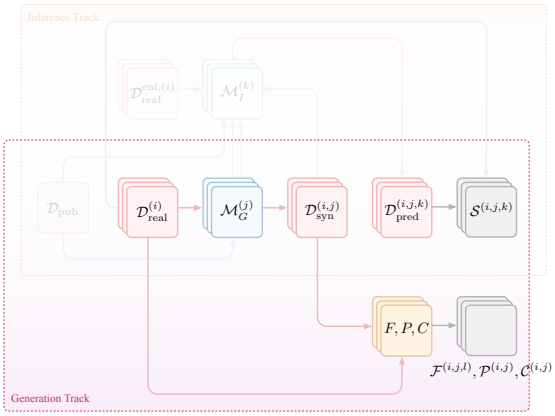
This challenge introduced a new dataset. AmsterdamUMCdb was developed and released by Amsterdam UMC in the Netherlands and the European Society of Intensive Care Medicine (ESICM). It is the first freely accessible comprehensive and high resolution European intensive care database. It is also first to have addressed compliance with General Data Protection Regulation (GDPR, EU 2016/679) using an extensive risk-based de-identification approach. However, both ESICM and Amsterdam UMC aim to continuously evaluate and if necessary improve privacy while maintaining usability.

AmsterdamUMCdb contains real data from critically ill patients from a mixed surgical-medical tertiary referral centre for intensive care medicine with up to 32 intensive care beds and up to 10 high dependency beds. It was released in early 2020 as comma separated value files, with a total uncompressed size of approximately 78 GB. Access may be requested through https://www.amsterdammedicaldatascience.nl and downloaded from https://doi.org/10.17026/dans-22u-f8vd. Detailed descriptions of the data schema and sample code to interact with the data are available on the AmsterdamUMCdb GitHub repository at https://github.com/AmsterdamUMC/AmsterdamUMCdb.

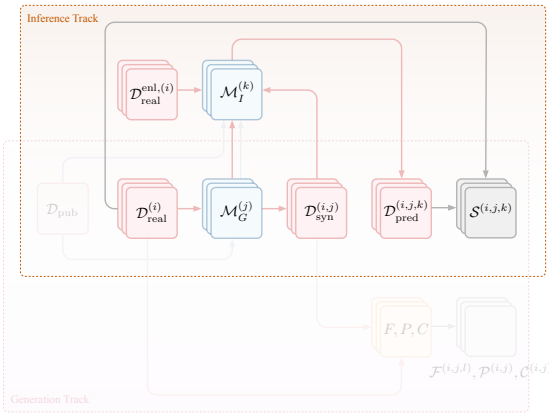
AmsterdamUMCdb contains approximately 1 billion clinical data points related to 23,106 admissions of 20,109 unique patients between 2003 and 2016. The released data points include patient monitor and life support device data, laboratory measurements, clinical observations and scores, medical procedures and tasks, medication, fluid balance, diagnosis groups and clinical patient outcomes. Data granularity depends on the type of data and admission year, but is up to 1 value every minute for data from patient monitor and life support devices. The data is much richer and granular than those in other well known freely available intensive care databases, such as MIMIC and is comprised of patients with higher illness acuity than is found in US datasets.



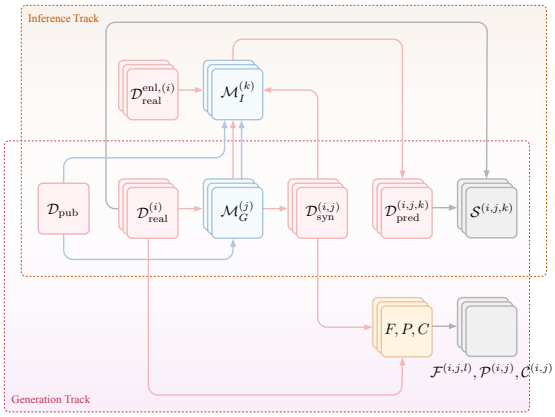
(*a*) **Participants and Submissions**. Each generation-track participant $j \in \{1, ..., N_G\}$ receives an initial $\mathcal{D}_{pub}$ for algorithm development; they each submit $\mathcal{M}_G^j$. Each inference-track participant $k \in \{1, ..., N_I\}$ similarly receives the initial $\mathcal{D}_{pub}$ for algorithm development, and moreover has access to all $\mathcal{M}_G^j, j \in \{1, ..., N_G\}$; they each submit $\mathcal{M}_I^i$.



(*b*) **Generation Track Evaluation**. We sample N subsets $\mathcal{D}^i_{\text{real}}$ from $\mathcal{D}_{priv}$, $i \in \{1, ..., N\}$, each of which is used to train $\mathcal{M}_G^j$, outputting $\mathcal{D}^{i,j}_{\text{syn}}$. Each $\mathcal{D}^i_{\text{real}}$ is also randomly enlarged to form $\mathcal{D}^{\text{enl},i}_{\text{real}}$. Submissions qualify if the degradation from TRTR to TSTR satisfies Equations 4–6. Once qualified, the overall score for each j is given by Equation 7.



(*c*) **Inference Track Evaluation**. We run each inference algorithm $\mathcal{M}_I^i$ on the tuple $(\mathcal{M}_G^j, \mathcal{D}^{\text{enl},i}_{\text{real}}, \mathcal{D}^{i,j}_{\text{syn}})$, which outputs predictions $\mathcal{D}^{(i,j,k)}_{\text{pred}}$ that aim to approximate $\mathcal{D}^i_{\text{real}}$. The score for each k is given by Equation 3.



(*d*) **Complete Competition Setup**. Blue lines denote the tasks of participants in submitting algorithms (blue boxes). Red lines denote the flow of input and output datasets (red boxes) to and from algorithms. Gray lines denote the computation of performance scores (gray boxes). The orange box denotes a collection of standard prediction tasks.

Figure 1: Schematics and descriptions for the mechanics of submissions and evaluations.

An accessible competition In order to make this first competition more accessible, we downsampled the dataset, selecting a subset of features and of time points in order to ensure that the competition was accessible to a variety of teams, but also to ensure that the computational complexity of hider algorithms was kept low. In particular, because hider teams were required to submit algorithms rather than trained models, using the full dataset would have been computationally infeasible.

4. Results

In total we received 23 hider submissions and 12 seeker submissions, from a total of 13 different teams. Of these, 9 and 8 were disqualified, respectively (e.g. for not providing documentation for their code, or being duplicate submissions). Of the 14 remaining hider submissions, only 4 passed the quality threshold outlined in Section 2.3.1. Table 1 contains the re-identification scores of these 4 hiders against the 6 (4 submitted + 2 baseline) seekers, with only the best submission from each team included.

Team Name	Mikan	k-NN*	GenSynDS	realHider	iCAIRD	BP*
Mikan1	0.5029	0.5017	**0.5036**	0.5000	0.4982	0.4992
Golden_Fleece1	**0.5037**	0.5019	0.5027	0.5000	0.5000	0.4940
Mikan2	0.4947	**0.5066**	0.4997	0.5000	0.4945	0.5006
Golden_Fleece2	**0.8131**	0.7508	0.7216	0.5000	0.4969	0.4944
Mean seeker score	0.5777	0.5650	0.5553	0.5000	0.4992	0.4950

Table 1: Results (* indicates a baseline algorithm and **bold** denotes the best performing seeker against the given hider)

In both sides of the competition, team *Mikan* won, with an noise-injection based approach on the hider side and a kNN based approach on the seeker side. The Mikan kNN approach differs from the baseline kNN in its preprocessing of the data. While the baseline kNN imputes missing data using medians, the Mikan submission imputes the missing data as zeros. Though we see that, overall, seeker scores are very close to 0.5 for all submissions except Golden_Fleece2, for which k-NN based scores are very high. If the Golden_Fleece2 hider submission is excluded, then all seeker scores except for the baseline k-NN, Golden_Fleece and realHider, drop below 0.5. We do not feel that these results are particularly meaningful, which will be discussed in the following section.

5. Lessons Learned

Structure Participants were given around 4 months to submit algorithms to both sides of the track. Both tracks were open for the full window. It was originally the hope that the two-tracked nature of the competition would allow for an evolving competition in which seekers would be able to target their algorithms to specific hiders. Unfortunately, because of this, hiders were not incentivised to submit their algorithms early (in fact the opposite - the later a hider was submitted, the less time the seekers have to hack it). This meant that seekers were not able to target specific hiders as was the original hope. In future iterations

of the competition, an incentive is needed for the hiders to submit early. At the very least, hider submissions should be closed before seeker submissions, to allow seekers time to "hack" the final hider submissions. Incentivising early submission also has the added benefit of ensuring bugs can be discovered early.

Hiders As noted in Section 4, many hiders failed to pass the quality bar. We believe this is due to the complex nature of the dataset, and thus in the next iteration of the competition, a simpler (perhaps more granular) dataset could be used, to allow the competition to really evaluate privacy preserving methods (instead, this competition mostly told us that generating high quality time-series data is hard). Moreover, the quality bar itself needs to be rethinked, with a more robust measure used, such as in Alaa et al. (2021).

Seekers As noted above, seeker submissions were originally intended to be able to run against specific hiders if they so wished. As we see in Section 4, no seeker algorithm performed well, with most having an almost 50% accuracy, equivalent to random guessing. This could, in part, be due to the complexity of the time-series dataset, and the lack of ability to target specific hiders.

6. Conclusion

Ultimately, the competition has highlighted the need not only for high quality synthetic data generation methods but also good, robust metrics for evaluating such data. It is our hope that a future competition involving such metrics, alongside a slightly revamped design, will create a strong platform through which we can evaluate existing ideas within the realm of private synthetic data generation.

Acknowledgments

We thank the Office for Naval Research (ONR), Alzheimer's Research UK and EPSRC for funding this research. We thank Microsoft Research for providing prizes and compute for the competition.

References

Ahmed M Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. *arXiv preprint arXiv:2102.08921*, 2021.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Measuring the quality of synthetic data for use in competitions. *arXiv preprint arXiv:1806.11345*, 2018.

N. Li, T. Li, and S. Venkitasubramaniam. t-closeness: Privacy beyond k-anonymity and l-diversity. *IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):1–52, 2007.

Olof Mogren. C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*, 2016.

W. N. Price and I. G. Cohen. Privacy in the age of medical big data. *Nat. Med.*, 25(1): 37–43, 01 2019.

Giorgia Ramponi, Pavlos Protopapas, Marco Brambilla, and Ryan Janssen. T-cgan: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling. *arXiv preprint arXiv:1811.08295*, 2018.

H. Shah. The DeepMind debacle demands dialogue on data. *Nature*, 547(7663):259, 07 2017.

L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 5509–5519, 2019.

Jinsung Yoon, Lydia N Drumright, and Mihaela van der Schaar. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE Journal of Biomedical and Health Informatics*, 2020.