BIG IDEAS

Synthetic data: breaking the data logjam in machine learning for healthcare

Mihaela van der Schaar Nick Maxfield September 7, 2020 20 min read

In this post, I will:
– Describe the problems surrounding access to, and use of, healthcare data
– Propose synthetic data as a solution to these problems
– Outline some of the synthetic data tools our lab has been developing
– Discuss evaluation of synthetic data, and the meaning of "quality"
– Present a vision for a synthetic data "clearing house"
– Call on the machine learning and healthcare communities to get involved.

Machine learning has the potential to catalyze a complete transformation in healthcare, but researchers in our field are still hamstrung by a lack of access to high-quality data, which is the result of perfectly valid concerns regarding privacy.

is one of a small handful of groups cutting a path through this largely uncharted territory. This is a very complicated but uniquely important endeavor, which combines conceptual and technical challenges and has required considerable open-mindedness as we approach it: we have even needed to come up with new ways of understanding notions such as data quality.

First, a quick disclaimer…

At this point, I should clarify that I do not plan to wade into debates surrounding the respective strengths or weaknesses of the well-known and much-discussed privacy preservation definitions and methods.

For our purposes, only three facts matter: 1) no universally accepted and quantifiable definition of identifiability exists, 2) it is the data guardians, not the data users, who set the terms for providing data, and they have all sorts of different requirements, and 3) regulatory efforts such as GDPR and HIPAA cannot provide adequate definitions, safeguards, or reassurance (for a discussion of why this is the case, see our March 2020 IEEE paper here).

While terms such as differential privacy may appear in the discussion below, please bear in mind that to us such notions simply represent potential "requirements" that may or may not need to be met, depending on the data guardian and the purpose of use of the data. Our lab's aim is to develop tools that can adaptably meet the full range of likely needs and requirements. This requires us to take a range of concepts related to privacy and data fidelity into account, even if we are agnostic about their merits.

Machine learning for healthcare: a delicate risk-benefit balance

As with many areas of research, machine learning for healthcare involves a balance between risks and benefits. On one hand, the information contained in electronic health records is inherently sensitive, and abuse of such information could cause great harm. On the other hand, the ability to make use of electronic health records (and other valuable information such as clinical trials, registry data, and so forth) for entire populations could completely transform healthcare research and delivery, driving life-saving insights and making new and powerful connections that we are currently unable to see. In fact, it can be argued that we are ethically bound to seek ways to use all available technology and advancements to improve healthcare—especially as demand is growing, but the availability of healthcare is not. In that sense, using data to empower machine learning would enable treatment of more patients, more efficiently, with better outcomes, while using the same resources.

As things stand, the risk-benefit balance is still perceived by many to be substantially lopsided, and data guardians are (understandably) reluctant to share patient data with researchers. The

through ad-hoc agreements with specific collaborators, or a very small handful of freely accessible databases.

This is particularly inhibitive in the field of machine learning for healthcare: access to data is the lifeblood of our research, yet we cannot conjure fresh data at will through real-world experimentation—unlike how autonomous car developers can, for example, generate valuable new data by sending cars out to drive more miles.

In a sense, then, our situation is not dissimilar in nature to the circumstances that prompted the creation of ImageNet over 10 years ago: we need to be able to retrieve, organize, and harness valuable data, but we are having trouble doing so. **We need our own equivalent of an "ImageNet moment" that catapults machine learning for healthcare forward by providing common datasets and frameworks.** In our case, the situation is further compounded by the additional need for a mechanism to ensure that these data are collected safely. **Our research task, therefore, is to identify the "sweet spot" in the risk-benefit balance, where high-quality yet sufficiently private data can be released.**

I believe the key to breaking this data logjam and energizing our research could lie in synthetic data. Synthetic data is an extremely promising but as-yet underexplored area that our lab has been working on over the last few years, and I would like to use this post to introduce some promising new approaches and present a vision for the way forward. This is quite an open-ended area of research, and my hope is that others will engage with us to solve these problems together.

Synthetic data: solutions and opportunities

There are several types of synthetic data, but the term essentially refers to the generation of artificial data with the aim of reproducing the statistical properties of an original dataset. These data can be either partially or fully synthetic, with the former containing generated data alongside original data, and the latter composed exclusively of generated data.

Synthetic datasets can be used to great effect across a range of applications within machine learning (and beyond). If the purpose of sharing a dataset is to develop and validate machine learning methods for a particular task (e.g. prognostic risk scoring), real data is not necessary; it would suffice to have a synthetic dataset that is sufficiently like the real data. Generating synthetic patient records based on real patient records can, therefore, be an alternative way of providing machine learning researchers with the data that they need to be able to develop appropriate methods for the task at hand, while avoiding sharing sensitive patient information. This could dramatically swing the balance between risks and benefits in favor of the latter.

created specifically for ICU admission prediction, for clinical trials, for estimating treatment effects, and for time-series data (to name a few examples).

I should also point out that the inherently shareable nature of synthetic data solves the problem of reproducibility of research: at present we often cannot provide training datasets to third parties who wish to verify our models, but using synthetic data would render this a non-issue.

Synthetic data for healthcare also comes with a complex array of challenges. There is no consensus, for example, regarding how to define or measure the quality of synthetic data (though some of our work addresses this, as outlined below). Data guardians also have a range of different requirements and expectations regarding privacy preservation, and these must all be accommodated, which means our solutions must be built to satisfy an array of different potential requirements. Lastly, the very advantage of being able to create different data types creates as many technical problems as opportunities (for example, "good" synthetic data will have different characteristics when developed for a time series setting, as opposed to a static setting).

All of this makes for a complicated and diverse (but also fascinating) research agenda. As pathfinders in this new area, our lab has developed a number of approaches to generating and evaluating synthetic data, some of which are outlined below.

Tools for generating synthetic data

First, it's worth bearing in mind that synthetic data is a term denoting the outcome of a process (data generation), rather than the process itself. There are many techniques and approaches that can be applied to actually generating synthetic data, some of which are outlined below.

Generating realistic synthetic data is challenging because patient records are often high-dimensional and come from complex distributions. Moreover, there will often be a relatively small number of unique observations (for example, patients with a rare disease or those who are outliers) and it is therefore even harder to accurately estimate the complex, high-dimensional distribution of these data without replicating the individual.

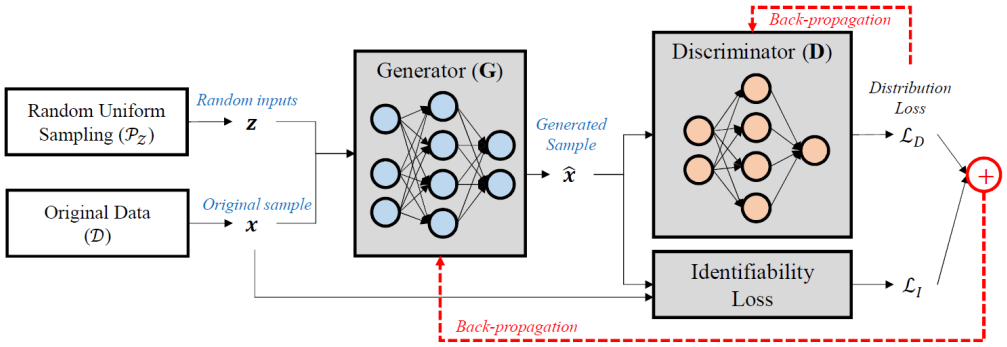
One particular area in which our lab has had considerable success is generative adversarial networks (GANs). GANs provide a promising framework for simulating complex distributions, and are already used in a range of other areas of application, such as synthetic image generation and image translation. The defining feature of the GAN framework is the existence of a generator and discriminator, trained in an adversarial fashion against each other. The

samples are the synthetic ones. This framework can be formulated as a minimax game and at the optimal point of this game, generated samples follow the real data distribution.

In 2019, our lab developed PATE-GAN, a technique we presented in a paper for ICLR. PATE-GAN is based on a modified version of the Private Aggregation of Teacher Ensembles (PATE), which provides a mechanism for classification by training multiple teacher models on disjoint partitions of the data; to classify a new sample using PATE, each teacher's output is evaluated on the sample and then all outputs are aggregated with noise.

PATE-GAN modifies the training procedure of the GAN's discriminator to be differentially private by using a modified version of PATE. Since the Post-Processing Theorem guarantees that the GAN generator—which is trained only using the differentially private discriminator—will also be differentially private and thus so will the synthetic data it generates. In essence, we can tightly bound the influence of any individual sample on the model, thereby resulting in tight differential privacy guarantees and thus improved performance over models with the same guarantees (although "performance" is a loaded term, as discussed below!).

Following PATE-GAN, in 2020 our lab developed a new data synthesis model, ADS-GAN, in collaboration with Lydia Drumright from the Centre for Cambridge Clinical Informatics. ADS-GAN (anonymization through data synthesis using generative adversarial networks), was introduced in a paper for the IEEE Journal of Biomedical and Health Informatics. It modifies the conditional GAN framework where the generator and discriminator are given additional inputs in the form of (the values of) a set of conditioning variables.



Block diagram of ADS-GAN. The generator uses original sample (x) and random vector (z) to generate a sample. The summation of distribution loss and identifiability loss is back-propagated to the generator. Both the generator and discriminator are implemented with multi-layer perceptrons.

generates all components based on these. We do this to improve the quality of the (fully) synthetic data while also ensuring that no combination of features could readily reveal a patient's identity. Within this model, the user can define their threshold of acceptable identifiability. This allows the model operator to weigh the risk benefit ratio for any given problem.

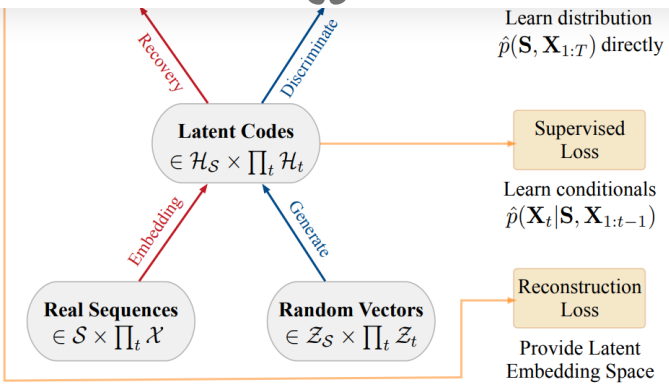
Our work around ADS-GAN also allowed us to move away from existing notions of identifiability, and instead define identifiability based on the probability of re-identification given the combination of all data on any individual patient.

We based this on the notion that synthetic patient observations should be "different enough" from the *original* patient observations in order to deidentify the original patient observations. We defined "different enough" as equivalent to the extent to which two different observations in the original dataset are "different enough" because they are "different" patients. Therefore, we use the "minimum" (closest) distance between the original observations as the measure for "different enough" between the synthetic and original data. Whoever is working with the model can set their own standards for "different enough," so that mildly sensitive data do not require as much noise, whereas more noise can be added to highly sensitive data.

(I should note again that we do not mean to suggest that this approach is superior to, or can replace, differential privacy or similar definitions of privacy. It's simply another option we can offer, depending on the requirements of the data guardians and intended application.)

The techniques described above have been shown to be effective in a static setting, but working with time series data is an essential aspect of machine learning for healthcare. However, the temporal setting poses a unique challenge, since models are not only tasked with capturing the distributions of features within each time point, but should also capture the potentially complex dynamics of those variables across time.

Existing methods do not adequately attend to the temporal correlations unique to time-series data. At the same time, supervised models for sequence prediction—which allow finer control over network dynamics—are inherently deterministic. This led our lab to develop TimeGAN, a generative model for time-series data, which we presented in a paper for NeurIPS 2019. TimeGAN straddles the intersection of multiple strands of research, combining themes from autoregressive models for sequence prediction, GAN-based methods for sequence generation, and time-series representation learning.



Block diagram of component functions and objectives.

Since TimeGAN is trained adversarially and jointly via a learned embedding space with both supervised and unsupervised losses, it offers both the flexibility of unsupervised GAN frameworks and the control afforded by supervised training in autoregressive models. There's more to TimeGAN than included above, and I'd definitely recommend that anyone interest in learning more have a look at our 2019 paper.

So far, I have focused on GANs as a group of tools that are particularly effective when applied to the task of generating synthetic data. There are, in fact, a variety of approaches that can likely be applied to the same task, such as the attentive state-space model of disease progression, which we introduced in a paper for NeurIPS 2019. This model uses an attention mechanism to create "memoryful" dynamics, whereby attention weights determine the dependence of future disease states on past medical history. Unlike GANs, the attentive state-space model attempts to capture the data distribution by explicitly modeling the physical process underlying disease progression with latent variables that correspond to interpretable clinical states. Through this modeling approach, we can generate not only observed clinical variables, but also trajectories of the underlying (unobserved) disease progression states in an unsupervised fashion. Because this likelihood-based model does not "memorize" the data trajectory for any specific patient, it has the advantage of being naturally privacy-preserving.

Our hope is that techniques such as the ones I've highlighted above could be used as part of a safe, legal, and ethical solution to open data sharing of datasets such as electronic health records, which would support the advancement of AI and machine learning in medicine. I will go into more detail on this later, but before doing so I would like to touch on another key issue related to synthetic data: quality.

How should synthetic data be evaluated?

help achieve this aim. I should, however, point out that we must still address some absolutely crucial (but somewhat abstract) questions about data quality: how do we tell whether a generated dataset faithfully reflects the real data, and what are the factors that determine the utility of a generated dataset for a specific purpose?

It might be tempting to assume that all that is required of synthetic data is "realism," but that in itself is a notion that defies a single definition. We are not, after all, simply photocopying a document and comparing it with the original. For example, synthetic data must not be compared on a feature-by-feature basis with the real data to check that the 1-dimensional distributions are similar; rather, it must be compared in terms of the joint distribution of features. But once we're considering the joint distribution of features, we also need to account for the fact that our requirements will depend entirely on the intended usage of the synthetic dataset in question.

In the healthcare setting, we will need synthetic data for predictions, survival analysis, clinical trials, causal inference, decision-making, competitions, and more. For each of these needs, specific types of synthetic data will be necessary, and these will all come with different required performance metrics, quality requirements, and even potentially privacy requirements. For instance, drug discovery requires that the effects of treatments in the real data are the same as in the synthetic data, but existing evaluation methods for synthetic data do not assess this.

For example, some use cases might benefit from a synthetic data generation method that involves training a machine learning model on the synthetic data and then testing on the real data. In such a case, if we were to see a high predictive performance on the real data for models that were trained on synthetic data, we can infer that the synthetic data have captured the relationship between features and labels well. Moreover, synthetic data that do well in this setting can be used to train models without ever seeing the real data. This is an important metric when the synthetic data will be used to train models to be deployed on real data, but is not always the most important metric.

When we consider synthetic data for use in competitions, by contrast, we would need synthetic data that allow researchers to do meaningful comparisons on the synthetic data. In this setting, the researchers will only be able to use the synthetic data as both the training and testing set (as they do not have access to the real data), and will need to develop their algorithms using results on the synthetic data.

In a 2018 paper, we explored a new key characteristic for assessing the utility of synthetic data for machine learning researchers in such a scenario: the similarity between the relative performance of two algorithms (trained and tested) on the synthetic dataset and their relative

on the synthetic data. This allows researchers to use the synthetic data to choose the best method(s) to try on the real data (or rather to give to the data-holder to try on the real data).

This is a crucial distinction because, rather than testing all possible algorithms simultaneously and selecting the best, a machine learning researcher will develop an algorithm over time by comparing a small set of algorithms, selecting the best, and then comparing the best within another small set of algorithms. It is therefore important that at each stage of this process, the best algorithm is selected. This means that comparisons between any two algorithms on the synthetic data should be similar to comparisons of the same two algorithms on the real data.

We called this approach to measuring the quality of synthetic data Synthetic Ranking Agreement (SRA). The SRA can be thought of as the (empirical) probability of a comparison on the synthetic data being ”correct” (i.e. the same as the comparison would be on the real data). A particularly interesting property of SRA is that it does not necessarily require the synthetic data to be distributed the same as the real data for a high SRA score to be achieved. This can be useful when we consider the implications this has for privacy, where training synthetic data generation models to be too similar to the real data can lead to concerns.

As outlined above, our own research covers a few potential approaches to evaluating the quality of synthetic data, but it is clear that substantial new research is needed, and must be conducted at scale.

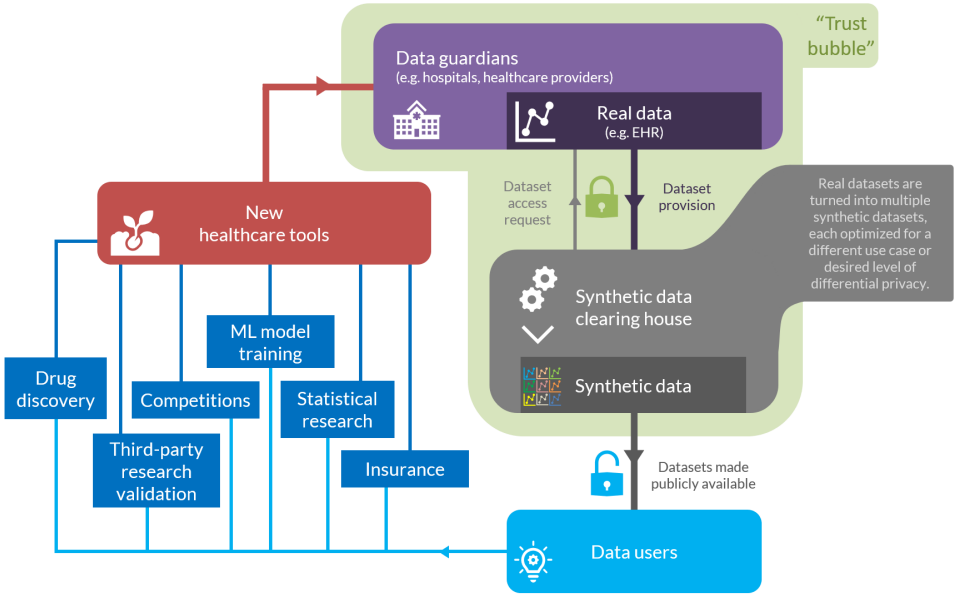
A “clearing house” for synthetic data

I would now like to break from my discussion of existing approaches, and talk about what we could achieve in the near future.

There are many different stakeholders within healthcare, but (on this topic) we can broadly divide them into two groups: data guardians, and data users. Data users are in dire need of high-quality data to apply to their research, whether related to statistics, machine learning model development, or any other number of cases. The vast majority of data guardians are (justifiably) unlikely to trust the vast majority of data users—despite the enormous potential societal benefits of doing so.

To kick-start this virtuous cycle, we need trust. In my opinion, the best way to generate that trust is through the creation of a commonly recognized body that handles sensitive data and generates synthetic data. I envisage this, in its broad strokes, as something akin to a clearing house: an entity that reduces counterparty risk and instills confidence in transactions by serving as a trusted (and supervised) intermediary.

converted into multiple versions of synthetic datasets, with different versions designed for different privacy requirements or usage cases. Data users would be able to obtain the synthetic data with relatively low barriers to entry.



The advantages of such an approach are considerable: data guardians would no longer need to worry about whether or not to trust individual data users, and could ultimately reap the benefits of new healthcare tools developed thanks to the data they provide; data users would be spared the exhausting (and often fruitless) process of seeking out individual data guardians and earning their trust, and instead could choose from an extremely broad array of high-quality and uniformly presented data; meanwhile, individuals in the real datasets would know that their own personally identifiable information would not leave the "trust bubble," and could also be offered the freedom of selecting the degree of information to be shared in the real datasets.

Implementing this vision together

As a group of machine learning experts, everyone in my lab knows the value—and scarcity—of high-quality data. My guess is that anyone who has read this far will understand this, too. The framework I've outlined above is just one of many potential solutions to a problem that urgently needs to be addressed. In sharing it, I hope to start a debate that one day coalesces into a consensus and, eventually, action.

If you'd like to learn more and get involved, please have a look at our publications on this topic and check out our hide-and-seek privacy challenge, which is part of the NeurIPS 2020

please, get in touch!

Related paper

Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods

James Jordon, Alan Wilson, Mihaela van der Schaar

Abstract

Synthetic data

Mihaela van der Schaar

Mihaela van der Schaar is the John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine at the University of Cambridge, a Fellow at The Alan Turing Institute in London, and a Chancellor's Professor at UCLA.

Mihaela has received numerous awards, including the Oon Prize on Preventative Medicine from the University of Cambridge (2018), a National Science Foundation CAREER Award (2004), 3 IBM Faculty Awards, the IBM Exploratory Stream Analytics Innovation Award, the Philips Make a Difference Award and several best paper awards, including the IEEE Darlington Award.

In 2019, she was identified by National Endowment for Science, Technology and the Arts as the most-cited female AI researcher in the UK. She was also elected as a 2019 "Star in

theory, distributed systems, machine learning and AI.

Mihaela's research focus is on machine learning, AI and operations research for healthcare and medicine.

VIEW ALL POSTS

Nick Maxfield

Nick oversees the van der Schaar Lab's communications, including media relations, content creation, and maintenance of the lab's online presence.

Nick studied Japanese (BA Hons.) at the University of Oxford, graduating in 2012. Nick previously worked in HQ communications roles at Toyota (2013-2016) and Nissan (2016-2020).

Given his humanities/languages background and experience in communications, Nick is well-positioned to highlight and explain the real-world impact of research that can often be quite esoteric. Thankfully, he is comfortable asking almost endless questions in order to understand a topic.

VIEW ALL POSTS

Revolutionizing healthcare: an invitation to clinical professionals everywhere

AutoML: powering the new human-machine learning ecosystem

You may also like

EVENTS VIDEO

Our fifth Inspiration Exchange engagement session took place on January 26 and was attended by over 100 AI and machine learning students.

Nick Maxfield Mihaela van der Schaar 6 hours ago

BIG IDEAS

Machine learning for healthcare: Towards a unifying framework

This post proposes a framework for machine learning for healthcare, based on past work and conversations with clinicians over many...

Mihaela van der Schaar November 8, 2020

BIG IDEAS EVENTS

Revolutionizing healthcare: an invitation to clinical professionals everywhere

In one week, we will hold our first "Revolutionizing Healthcare" discussion on how machine learning can change...

Mihaela van der Schaar Nick Maxfield September 22, 2020

View posts about…

Augmented MD AutoML

Conference COVID-19

Cystic Fibrosis Deep learning

Ensemble learning

Feature selection ICLR

ICME ICML Interpretability

Journal Missing data imputation

MLHC Model NeurIPS

Paper Personalized medicine

Pharma Policy Impact Predictor

Reinforcement learning

Software Statistics

Synthetic data

Time series analysis

Transfer learning Transplant

Upcoming events

29
JAN

RE•WORK DEEP LEARNING 2.0 VIRTUAL SUMMIT PRESENTATION