Digital Professionalism in Health and Care: Developing the Workforce, Building the Future
P. Scott et al. (Eds.)

doi:10.3233/SHTI220925

An Effective Approach for Extending Medical Data to the Cloud Through Synthetic Data Generation for Educational Environments

Alan BOYCE[a,1], Michael DACEY[b] and Tim BASHFORD[b]
[a]*National Data Resource (NDR), Digital Health and Care Wales (DHCW), Wales, United Kingdom*
[b]*School of Computing, University of Wales Trinity Saint David*

Abstract. When taking advantage of technology, healthcare is often met with considerably more barriers to entry than business. Cloud platforms can offer great benefits such as scalability, reduced cost and the ability to effortlessly collaborate across services, and indeed, across the world [6] yet healthcare has been slow to take advantage of these gains. This paper explores the challenges faced by healthcare, how using synthetic data can avoid the initial information governance barriers, provide the experience to effectively evaluate cloud platforms, enable effective research collaboration with education and industry, and support the digital transformation journey.

Keywords. Cloud first, Cloud evaluation, Education, Simulation, Medical data, Healthcare, Synthetic data, Data generation, Digital transformation, Information governance, Security, Privacy, Wales, United Kingdom

1. Introduction

When exploring Cloud technologies, we are met with many challenges. Firstly, healthcare systems are comprised of many owners, services, applications and independent infrastructures. This makes the move to the cloud difficult to initiate as ownership needs to be defined upfront in order to set up effective billing structures. It also makes selecting an initial use case problematic as you must decide between a realistic cross-section that will require many stakeholders or single stakeholder project that fails to contain the depth of services to challenge the platform.

Secondly, due to information governance and General Data Protection Regulation (GDPR), moving data to cloud platforms involves high levels of assurance, access management and security that your workforce may not yet have had the training or experience to be held responsible for. As healthcare contains many levels of Personally Identifiable Information (PII) and even anonymised or pseudo-anonymised data can still be identifiable due to small datasets, real-world data is often ruled out. Certainly, at the early stages there is a need to explore options in order to make informed decisions. This

[1] Corresponding Author, Alan Boyce; E-mail: alan.boyce@wales.nhs.uk.

often leads to proof of concepts with small generic artificial data sets or generic datasets that are not representative of real-world data. Furthermore, choosing subsets or generic demo data hides the complexities faced in transforming real-world data, how research spaces may wish to explore the datasets and at what stages data will need to be de-identified for different use cases. The challenge here is ensuring that the generated data is close enough to the original so that it is usable in practice.

Thirdly, due to the complexity and orchestration of multiple services, the cost of cloud computing is difficult to estimate upfront. When exploring the costs and suitability of the approach it can be tempting to use a small cross section or use case to validate the platform capability and estimate the costs involved. Unfortunately, this approach often excludes many of the nuances of the data and processes required to obtain and prepare the data for use. The issue is further compounded by PII that excludes real-world data from being used. This can lead to the evaluation datasets being a fraction in size and not representative of the range of information that needs to be consumed by the platform.

Another key requirement is to provide research collaborators in education and industry a cost-effective means to simulate a realistic infrastructure so the transition to work with real data is effective and efficient. Having synthetic data that fully represents the original source with no sensitive information provides an easy way to share data.

The health service sector traditionally approaches solutions wholly onsite. However, over recent years business has seen clear benefits when transitioning to a Cloud First approach. Many health services have been provided centrally in the past but in order to scale, they are distributing the delivery across many locations, units, communities and third parties. As a result, there is a need to capture and distribute the patient information across services whilst keeping data as close to real-time as possible to ensure the validity as they are transferred between specialised services. This has prompted many health services to explore transitioning to the cloud. NHS Wales has tackled this with a dedicated programme, our National Data Resource (NDR) team was tasked with a Cloud First approach to be used for all new medical data projects across Wales. Ideally the approach must be reusable in order to evaluate multiple cloud providers, although in practice successfully generating the data in one platform will allow the same generated data to be used across the others.

In addition to the technical complexities and challenges, there are also challenges in information governance and digital transformation. Most healthcare workforces have a strong background in traditional databases and application management. Therefore, many employees are new to cloud and a data engineering approach. They have numerous legacy systems to support and are under constant pressure to perform with drastically reduced budgets and resources. Training can provide the background and concepts required but fundamental changes to the ways of working are required in order to develop truly cross-functional teams.

Could using synthetic data that closely represents the source data better inform decision making? Could it support the digital transformation journey of the workforce by providing near real-world use cases for evaluation? Could it provide the right level of challenge to form cross-functional teams to develop key core concepts and experience, critical to future decision making? Could using synthetic data mitigate restrictions on the physical location and country boundaries that some data is bound to?

This paper explores synthetic data generation as a way to ease the transition to the cloud and its effectiveness as a tool to better decision making.

2. Methods

We began by considering the various tools and techniques available to generate synthetic data [3]:

1. Discovering patterns from imbalanced data can play an important role although it has its limits in high-dimensional data, such as images. Using a variational autoencoder (VAE) based synthetic data generation for imbalanced data can produce new samples that are similar to the original dataset [2].
2. Take a sensitive dataset as input and generate a structurally and statistically similar synthetic dataset, with strong privacy guarantees, as output [5].
3. Use deep learning techniques to capture the relationships among multiple features and then use the models to generate differentially private synthetic datasets [1].
4. Machine learning has made a significant impact in medicine and cancer research [4] and machine learning-based tools like SynSys can be effective in creating a model to generate synthetic data. A drawback of this technique is that it requires significant training time on real datasets which is often difficult due to the same privacy and governance issues we are trying to avoid by taking the synthetic data approach.
5. Additional techniques may be required for complex, multidimensional datasets such as image data. Such data is often limited due to a lack of data samples and privacy reasons. As a result, creating synthetic data for training models could also be considered [7]

Each approach was an involved process, so we needed to begin by selecting an initial starting point with a view to extend this in subsequent iterations. The criteria for selecting the subjects were as follows: As the purpose is to evaluate the cloud, the initial process should require minimal cloud configuration and knowledge. Transferring real data, even for the purposes of generating synthetic equivalents, will raise the same information governance and security concerns so transferring little or no real-world data to the cloud environment would be advantageous. As an evaluation and learning process it is important to keep new technologies and systems to a minimum.

As a first step, applying these criteria, only option 2 met the condition of minimal cloud configuration and knowledge as the others required extensive Machine Learning (ML) or Artificial Intelligence (AI) knowledge. Again, only option 2 met the condition of not requiring sensitive data to be sent to the cloud in order to generate it. Option 2 also gave a streamlined evaluation and learning process.

We have selected option 2 from the above list. This requires us to take sensitive data as input and generate closely matching data. Synthea was selected to generate synthetic data through custom configuration files and state machines. It was selected as it required the least upfront cloud knowledge, no datasets had to be transferred to the cloud environment, and it only requires Java to run.

Firstly, we explored the elements of synthetic data generation that were important to our use case. We examined the methods available to generate synthetic data and how customisable they were. We confirmed they were suitable for the various types of medical data we require. We ensured we could automate the process with minimal effort. We examined the synthetic data to ensure it was close to real-world data and schema.

We selected appropriate uses cases that would help to inform important decisions in the future. Finally, we ensured that the approach is configurable and repeatable for all leading cloud providers using IaC (Infrastructure as Code).

3. Results

By qualitative observation of working practices involving heuristic rules for generating suitable rules for testing at Digital Health and Care Wales (DHCW), I found the following results.

Senior Data Analysts in the informatics team reviewed the data to ensure that it met the range and broad statistical qualities of the source data. For the purpose of platform evaluation, it was observed that synthetic data offered a close match to the original source and could also contain examples of sensitive information that would normally need to be removed. As the data only represented real-world data, it was possible to share synthetic data between education, industry and services that would not otherwise have been possible.

It was apparent from the additional tasks completed with synthetic data that more real-world scenarios could be explored. A comparison of the anonymised and synthetic results showed that synthetic data offered a more accurate example of the data and enabled a wider range of services and use cases to be explored.

The majority of respondents felt it allowed the team to work cross-functionally at a level that was not overwhelming. Using synthetic data allowed the team to work with data that resembled the real-world and was familiar to them. It was found that less time was spent with governance and administration activities required for real-world data. In many cases, data analysts existing access was sufficient for the original data in the local environment and generated synthetic data was used in the cloud. It was possible to openly share and collaborate on datasets that included schema and meaningful content.

After developing with synthetic data, it was evident the team was more confident in decision making, was prepared to take more responsibility in decisions, and had real world experience to fall back on.

The initial findings are limited by the scope of the proof of concepts, the time allocated to this exploration phase and the credits available proceeding the procurement phase. Further research should be done to investigate the selected method and expand it to evaluate the effectiveness of the other valid approaches and their impact at each stage of the evaluation.

4. Conclusions

These findings suggest synthetic data generation is more effective than anonymised and generic datasets when used in proof of concepts and evaluation. Although alternative methods using ML and AI could potentially produce more detailed data, the approach requires the data to be transferred to the cloud environment and is therefore not suitable for the initial evaluation. However, these approaches could be considered in later stage evaluations where a secure environment that meets the additional criteria is created for this purpose.

The study has shown that synthetic data allowed more real-world scenarios to be explored. Real data required much of the data to be anonymised or pseudo-anonymised.

In some cases, large portions of the schema were removed completely. It was also possible to generate data to match the 5 V's of (velocity, volume, value, variety and veracity) and was equally suitable to streaming and bulk loading.

Using synthetic data allowed the team to work with data that resembled the real world and was familiar to them. They were able to spend less time with governance and administration activities required for real-world data and were able to openly share and collaborate on datasets that included schema and meaningful content.

We found the team was more confident making larger decisions and had real-world experience to fall back on due to the more involved use cases they were able to explore. Furthermore, we found that using synthetic data was an effective way of producing an educational environment to analyse cloud-based health care and analysis systems.

Using synthetic data, with no sensitive information, but without compromising the original data and schema would allow us to share a realistic environment with an industrial or educational healthcare partner. This could be useful for Secure Anonymised Information Linkage (SAIL) or Secure eResearch Platform (UK SeRP) where researchers could explore synthetic datasets before hackathons or research.

These findings, while preliminary, suggests that building small scale solutions together as a team can help develop the ways of working and increase the levels of understanding. Simulating synthetic data allows the teams to develop services in an almost identical approach to the real-world systems. It also creates a reusable pipeline for generating test data that accurately represents the distributions and key parameters of the populations involved. We propose using a combination of synthetic data generation processes to build cloud solutions that closely resembles the real-world, ease collaboration challenges, support reusable test data and educational/learning environments.

Acknowledgements

We would like to acknowledge DHCW, Amazon Web Services (AWS) and University of Wales Trinity Saint David (UWTSD) for supporting this research.

References

[1] Abay NC, Zhou Y, Kantarcioglu M, Thuraisingham B, Sweeney L. Privacy preserving synthetic data release using deep learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2018 Sep 10 (pp. 510-526). Springer, Cham.

[2] Dankar FK, Ibrahim M. Fake it till you make it: Guidelines for effective synthetic data generation. Applied Sciences. 2021 Feb 28;11(5):2158.

[3] El Emam K, Mosquera L, Hoptroff R. Practical synthetic data generation: balancing privacy and the broad availability of data. O'Reilly Media; 2020 May 19.

[4] Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC medical research methodology. 2020 Dec;20(1):1-40.

[5] Howe B, Stoyanovich J, Ping H, Herman B, Gee M. Synthetic data for social good. arXiv preprint arXiv:1710.08874. 2017 Oct 24.

[6] Kagadis GC, Kloukinas C, Moore K, Philbin J, Papadimitroulas P, Alexakos C, Nagy PG, Visvikis D, Hendee WR. Cloud computing in medical imaging. Medical physics. 2013 Jul;40(7):070901.

[7] Thambawita V, Salehi P, Sheshkal SA, Hicks SA, Hammer HL, Parasa S, Lange TD, Halvorsen P, Riegler MA. SinGAN-Seg: Synthetic training data generation for medical image segmentation. PloS one. 2022 May 2;17(5):e0267976.