Pre-print

The potential synergies between synthetic data and in silico trials

Puja Myles[*], **Johan Ordish**[*], **Allan Tucker**[†]

[*] Medicines and Healthcare Products Regulatory Agency, London, UK.

[†] Department of Computer Science, Brunel University London, London, UK.

E-mail: puja.myles@mhra.gov.uk

Abstract:

In silico trial methods promise to improve the path to market for both medicines and medical devices, targeting the development of products, reducing reliance on animal trials, and providing adjunct evidence to bolster regulatory submissions. In silico trial methods are only as good as the simulated data which underpins them, consequently, often the most difficult challenge when creating robust in silico models is the generation of simulated measurements or even virtual patients that are representative of real measurements and patients. This article digests the current state of the art for synthetic data and provides a number for suggestions for potential synergies to unlock the potential of in silico methods by exploiting synthetic data to model effects on a more diverse population. Nascent work on synthetic data has the potential to assist with these issues, so far proving to be much more than a robust privacy enhancing technology. Synthetic data could be defined as artificial data that mimic the properties and relationships in real data. Recent advances in synthetic data generation methodologies have allowed for the generation of high-fidelity synthetic data that are both statistically and clinically, indistinguishable from real patient data. Other experimental work has demonstrated that synthetic data generation methods can be used for selective sample boosting of underrepresented groups. This article will provide a brief outline of synthetic data generation approaches and discuss how evaluation frameworks developed to assess synthetic data fidelity and utility could be adapted to evaluate the similarity of virtual patients used for in silico trials, to real patients. The article will then discuss outstanding challenges and areas for further research that would advance both synthetic data generation methods and in silico trial methods. Finally, the article will also provide a perspective on what evidence will be required to facilitate wider acceptance of in silico trials for regulatory evaluation of medicines and medical devices, including implications for post marketing safety surveillance.

Keywords: in silico trials, synthetic data; medical devices

Introduction

In silico trial methods represent an opportunity to augment and streamline elements of the path to market for both medical devices and medicines. Broadly, some of the promise is that further use of such methods might reduce reliance on animal trials and bolster evidence that would otherwise be generated at risk to clinical trial participants.[1] Accordingly, so long as the evidence is robust, the more that can be mustered from modelling, the smoother the route to market will be and less risk will be borne by participants. Indeed, there is emerging consensus of in silico's potential, the trajectory being that these methods have a role in evidencing medicines and medical devices, the primary questions now being how and to what extent? However, this potential is all predicated on access to quality data. In silico models trained on poor data will themselves perform poorly; models trained on incomplete data will be incomplete.

As access to quality data is likely to be the foremost challenge in getting in silico trial methods into standard practice, it is necessary to consider methods that might both facilitate access to data and methods that might boost underrepresented subgroups within datasets. For instance, it is one question to consider to what extent in silico methods are appropriate for regulatory purposes for both medicinal products and medical devices. In this article we define synthetic data as well as the various approaches used to generate such data, then outline a framework to evaluate synthetic data, also considering potential synergies between synthetic data and in silico trial methods, and then finally consider both areas for future research and regulatory questions that require further investigation.

Defining synthetic data

Conceptually, synthetic data are artificial data that mimic the properties of and relationships in real data. The quality of synthetic data depends on the approach taken to synthetic data generation and is often described in terms of its "utility" or "fidelity." A synthetic dataset that captures complex inter-relationships between various data fields and the statistical properties of real data can be referred to as a "high-fidelity" synthetic dataset.[2] It would follow that a "high-fidelity" synthetic dataset should also have "high-utility" i.e., the capability to produce analysis results similar to the original data.[3]

Using the example of patient healthcare data, a high-fidelity synthetic dataset would be able to capture complex clinical relationships and be clinically indistinguishable from real patient data. The generation of high-utility synthetic data tends to be highly resource intensive given the present state of play and depending on the application for which synthetic data are required, it may be acceptable to use low or medium utility synthetic data.

While high-fidelity synthetic data could be used as a proxy for real data (including for complex multivariable analyses involving a range of machine learning algorithms) with a high degree of confidence, medium-fidelity data would only be suitable for simple analyses like proportions, summary statistics for single variables or cross-tabulations involving two variables. Low-fidelity synthetic data on the other hand, should only be used as a sample dataset that provides an understanding of the data types, data values, data formats, data

[1] Badano A. In silico imaging clinical trials: cheaper, faster, better, safer, and more scalable. *Trials*. 2021;22(1):64. doi: 10.1186/s13063-020-05002-w.
[2] Myles P, Ordish J, Branson R. Synthetic data and the innovation, assessment, and regulation of AI Medical devices. *RF Quarterly*. 2021; 1(2): 48-53.
[3] El Emam K, Mosquera L, Hoptroff R. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media; 2020.

structure and table relationships in the real data that it seeks to represent. In the context of in silico trials, high-fidelity synthetic data would be required.

Synthetic data generation approaches

Synthetic data generation methods can be broadly categorized into three groups: generating synthetic data based on statistical properties of real data; adding noise to real data; and using machine-learning techniques to generate synthetic data.[4]

Generating synthetic data based on statistical properties of real data

This approach relies on statistical properties of real data such as population distributions – for example, mean values, standard deviation, and value ranges for data fields such as blood cholesterol measurements or known prevalence of a disease in various subgroups. This approach is useful when the real data are difficult to access, or the distribution of events is highly imbalanced in the available real data sample. A key limitation of this approach is that, while each synthetic data field will have the statistical properties of real data, the complex relationship between data fields will be difficult to capture. Thus, this approach would generally yield low- or medium-fidelity synthetic data.

Adding noise to real data

This approach involves perturbation of some of the data fields in real data in different ways including substitution of real values with other realistic values, random shuffling of data values within a particular data field or application of a random numeric variance (for e.g., plus/minus 10% applied to all data values in a field such that the data distribution is preserved). Substitution of real values can also be approached by swapping data within a data field with another sample from the same distribution.[5] These techniques can be used to generate low- or medium-fidelity data.

Machine-learning techniques to generate synthetic data

Machine-learning techniques such as Hidden Markov models, Bayesian networks (BNs), and deep-learning approaches such as generative adversarial networks (GANs) can be used to learn patterns between different data fields in real data. The learned patterns are then used as an input for the synthetic data generator to yield synthetic data. These methods can be used to yield medium- or high-fidelity synthetic data because they are able to capture complex relationships between various data fields.

The actual choice of machine-learning algorithm is dependent on the specific requirements for synthetic data. For instance, when transparency is a key requirement, BN approaches are preferable to GAN-based approaches. Unpublished findings from the MHRA's synthetic data research team suggest that GAN-based approaches may perform better than BN approaches for numerical data fields and vice versa for categorical/nominal data fields. The BN approach included latent variable modelling to deal with missing values in the real data. Hidden Markov models on the other hand, have been particularly useful for time-series data and are able to take into account missing values in real longitudinal data.[6]

Evaluation framework for synthetic data

[4] Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence*. 2021;37(2):819-51.

[5] Patel N, Patel S. A study on data perturbation techniques in privacy preserving data mining. *International Research Journal of Engineering and Technology*. 2015; 2(9): 2120-2124.

[6] Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*. 2020;3(1):1-3.

Data utility measures are a good way to assess whether a synthetic dataset can justify the claims of being high-fidelity. One of the earlier papers considering evaluation of synthetic data, Snoke et al. (2018) outlined general and specific utility measures for synthetic data.[7] They defined general utility measures as summaries of differences between real and original data as opposed to specific measures of utility that focused on results from particular analyses. They suggest that when the intended purpose of the synthetic dataset is known, specific measures of utility may be more helpful but when the intended purpose is not known, general utility measures are more appropriate.

More recently, El Emam et al. (2020) describe three types of approaches to assessing the utility of synthetic data: workload-aware evaluations, generic data utility metrics and subjective assessments of data utility.[8] Workload-aware metrics consider which types of analyses are feasible using the synthetic data and by replicating analyses carried out in the real data using the synthetic data. Analyses can range from simple descriptive statistics to complex multivariable machine learning models. Subjective evaluations involve classification of a random mix of real and synthetic records by domain experts followed by an evaluation of the accuracy of that classification. Generic assessments include metrics like the distance between the original and transformed data; these assessments provide an assessment of fidelity with utility being inferred on this basis.

Distributions can be compared by visual examination of histograms or by using summary statistics like the Hellinger distance (a probabilistic measure between 0 and 1, where 0 indicates no difference between distributions) to measure the difference in distributions between each variable in the real and synthetic data. The median Hellinger distance across all variables should be close to 0 with very small variations, for a high-fidelity dataset. Bivariate and multivariate distance analyses typically involve correlation analyses

Our own experiments in synthetic data generation have used a combination of all three approaches described by El Emam et al. (2020), using generic assessments of fidelity like the univariate, bivariate and multivariate distances between variables.[9] We used the Kolmogorov-Smirnov (KS) test to determine any differences in the univariate distance between the synthetic and real datasets and nonmetric multidimensional scaling (NMDS) to assess multivariate distance. We also undertook a subjective evaluation whereby two independent medical assessors reviewed a sample (n = 100) containing randomly selected records for equal number of synthetic and real patients with the aim of categorising them as synthetic or real based on the clinical characteristics. Finally, we compared the real and synthetic datasets by using stacked ensembles including six different machine learning algorithms [least absolute shrinkage and selection operator (LASSO), classification and regression training (CARET), extremely randomised trees, feed-forward neural networks, non-negative least squares and random forest] to predict cardiovascular disease risk for a more rigorous test of fidelity. This approach shares a similar philosophy to the ‘all models test’ approach proposed by El Emam et al. (2020) where all possible models are examined as it is not known a priori what an actual analyst would want to do with the dataset. Based on these evaluations, we posited that our approach to synthetic data generation using Bayesian

[7] Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2018;181(3):663-88.
[8] El Emam K, Mosquera L, Hoptroff R. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media; 2020.
[9] Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence*. 2021;37(2):819-51.

networks incorporating latent variables to learn the distributions and relationships in the real data, yielded high fidelity synthetic data.[10]

Such an evaluation framework could also apply to virtual patient cohorts employed in the context of in silico trials by providing a meaningful comparison to real patient cohorts. This would also be applicable to some degree to virtual patient cohorts that include boosted characteristics or simulated values to address missing data gaps in real data, though further work is needed in this area.

Potential synergies between synthetic data and in silico trials

High-fidelity synthetic data capture many of the complexities of real patient data. It offers the ability to infer the effects of medical interventions on a diverse population if generated using models of large national datasets. This has been possible for our approach because the CPRD database covers underlying health conditions of many different subpopulations within the UK, incorporating effects of, for example, age, ethnicity and regional disparity. Our approach to synthetic data generation means that we can condition our sampling of synthetic patients on evidence. For example, we may want to sample patients who suffer from a particular condition or from a specific demographic. This means that we can control for outcomes of virtual clinical trials to explore the effects more widely.[11] However, the utility of synthetic data can be limited by reliance on high-quality secondary data. That is, data collected for reasons other than simulating the effects of interventions.[12] This can potentially result in models that only reflect what has been measured in a population in the past and will not include effects of previously unseen interventions. One method to deal with this can be the linking of synthetic data to simulation approaches for in silico trials. These involve trying to capture the effects of interventions on individuals directly by using mechanistic models and should be able to complement synthetic data methods by identifying proxies of interventions in the synthetic data features which will subsequently allow inference on the wider impacts across different health outcomes of the population.[13]

Areas for further research

Linking high-fidelity synthetic data to virtual / in silico clinical trial data offers great potential. However, this research is still in its infancy and the identification of suitable proxies for linking the two data sources will be key to its success. There will need to be a full exploration of bias in clinical trials using appropriate metrics on sub-populations. We have begun this process on synthetic data by using boosting methods applied to certain sub-populations that are identified as under-represented based upon model performance metrics.[14] We are undertaking further experiments to determine whether such boosting is informative.. Furthermore, research is required to fully understand under what circumstances in silico

[10] Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*. 2020;3(1):1-3.
[11] Samei E, Kinahan P, Nishikawa RM, Maidment A. Virtual clinical trials: Why and what (special section guest editorial). *Journal of Medical Imaging*. 2020;7(4).
[12] Boslaugh S. *Secondary data sources for public health: A practical guide*. Cambridge University Press; 2007.
[13] Sarrami-Foroushani A, Lassila T, MacRaild M, Asquith J, Roes KC, Byrne JV, Frangi AF. In-silico trial of intracranial flow diverters replicates and expands insights from conventional clinical trials. *Nature Communications*. 2021;12(1):1-2.
[14] Draghi B, Wang Z, Myles P, Tucker A. Bayesboost: Identifying and handling bias using synthetic data generators. *AIIM-D-22-00323*; 2022.

models developed on synthetic data would be acceptable for regulatory purposes versus real data and what requirements would attach to those models.

Regulatory perspective

There is growing acceptance that in silico modelling has a role in evidencing medicines and medical devices. Synthetic data has the potential to accelerate in silico modelling. At minimum, we suggest that in silico models would have to demonstrate fidelity to their real data counterparts or comparative performance versus models trained on real data. As described above, the methodology for evaluating synthetic data is still nascent and developing. In the context of regulation, until the state of synthetic data crystalises, it is likely that the use of in silico modelling that typically requires some simulation akin to synthetic data will be stymied. It is therefore likely that the trajectory of further acceptance of in silico modelling will continue in the regulatory sphere, but further acceptance and standardisation of synthetic data will be necessary to accelerate acceptance.

Conclusion

The benefits of in silico modelling are plain: so long as the models are accurate, these methods provide a useful adjunct data that should increasingly make a contribution to the evidence base of medical devices and medicines. Synthetic data constitutes a complementary set of methods that presents obvious synergies to unlock and bolster datasets that underpin in silico models. Consequently, if the acceptance of synthetic data is stymied so too will the development of in silico models. Nevertheless, as methods to assess synthetic data progress, the benefits that it provides may outweigh its risks, thereby driving its acceptance amongst regulators.