JACC: BASIC TO TRANSLATIONAL SCIENCE VOL. 3, NO. 5, 2018

EDITOR'S PAGE

Are Synthetic Data Derivatives the Future of Translational Medicine?

Randi Foraker, PhD, MA,[a] Douglas L. Mann, MD, *Editor-in-Chief, JACC: Basic to Translational Science*,[b] Philip R.O. Payne, PhD[a]

As noted in this Editor's Page previously, the rising cost of developing new cardiovascular therapies cannot be sustained in the long-term (1). Accordingly, there is a critical need for new methodologies that can improve the speed, efficiency, and success rate of efforts to develop new therapeutic strategies for cardiovascular disease (2). Although randomized clinical trials remain the gold standard to evaluate drug responsiveness, phase III clinical trials are costly due in part to the large numbers of patients that need to be enrolled and the long follow-up period needed to detect meaningful differences in survival or clinical outcomes (3). As an alternative to randomized clinical trials, clinical effectiveness studies can be conducted to evaluate drug responsiveness in diverse patient populations (4,5). Such pragmatic approaches to evaluate drug responsiveness can be randomized or nonrandomized. If validly conducted, such studies can provide decision-makers with evidence from patients who are representative of those presenting to a clinic for a particular problem, thus accelerating translation into general clinical practice.

Treatment decisions that must be made by clinicians include: *Of existing treatments, which is best for an individual patient; what is the best treatment approach for patients with certain medical conditions; and how does one treatment compare with other existing alternatives?* Ideally, clinicians would have the ability to query the electronic health record or another patient database for treatment efficacy from a population of similar patients in order to guide treatment decision-making for an individual patient (6). In the absence of these types of data or that of clinical trials, the quality of evidence available to answer these critical questions is frequently insufficient. Rarely are studies conducted to assess treatment effectiveness or patient outcomes in real-world practice settings, and often trials are not designed nor powered to evaluate the comparative effectiveness of treatments (7).

To fill this gap, data are needed, not only to know how best to treat individual patients, but also to develop and refine evidence-based treatment guidelines. Decision-makers in need of this information include policymakers, payers, health care organizations, clinicians, and patients. To precisely estimate effect sizes, researchers must have access to sufficiently large and representative datasets. Although data sharing is an option to increase the sample size of an eligible study population, many institutions lack the infrastructure and support to do so (8). As a result, there are few networks of investigators who are willing and able to share data at the necessary scale in order to study drug responsiveness. This is a critical obstacle to progress, as data re-use and data sharing are essential for multisite, generalizable insights.

Synthetic data derivatives offer one potential solution to the aforementioned problems (9). Synthetic datasets are generated from existing datasets and maintain the statistical properties of the original dataset. Importantly, rows of observations in synthetic datasets do not correspond to identifiable individuals (rows of data) from the original dataset. Thus, synthetic data derivatives are quantitatively identical to patient-derived datasets, yet cannot be linked to the individuals from whom the data were derived (9). Because synthetic data contain no protected health information, the datasets can be shared freely among investigators or those in industry, without raising patient privacy concerns. In addition,

From the [a]Institute for Informatics, Washington University School of Medicine, St. Louis, Missouri; and the [b]Center for Cardiovascular Research, Cardiovascular Division, Washington University School of Medicine, St. Louis, Missouri.

ISSN 2452-302X https://doi.org/10.1016/j.jacbts.2018.08.007

research conducted using synthetic derivatives does not require institutional review board approval.

Notably, data synthesis differs from the anonymization or de-identification of protected health information through the removal of identifiable data elements or their obfuscation (10). Alternative approaches to synthetic derivatives include establishing a data enclave with restricted access and data-sharing requirements, or limiting access to only data that are relevant to a specific research question (11). Each of these alternatives does not ensure data privacy, because de-identified data can be re-identified with linkage to another data source, and security and confidentiality breaches can occur even with limited access to protected systems.

Using a data synthesis platform allows for the linkage of multiple sources of data before producing a synthetic derivative, and reduces data ownership concerns when combining data across organizational boundaries. Having the capability to combine datasets before synthesis results in a data product that provides a more comprehensive view of the patient, and facilitates the evaluation of factors related to drug responsiveness including those of health care quality and patient safety. For researchers, the ability to produce and share synthetic datasets can shorten the idea-to-insight time from years (as with expensive, lengthy clinical trials) to hours, and lessens legal and ethical barriers to data sharing. Not only does access to synthetic data allow for efficiencies in research, but the potential of synthetic data is great for saving time and money in drug development and responsiveness as well.

Can synthetic data be used to evaluate drug responsiveness? One of the major difficulties in developing new therapies relates to the inherent fragility of phase II trials. Because of cost constraints, the sample size of patients enrolled in early-phase trials is relatively small, and the number of drug doses that one can study feasibly is often limited. The size of phase II trials also restricts the range of endpoints that one can measure to gauge clinical effectiveness. Further, phase II trials are often performed in large academic medical centers that serve as tertiary and quaternary referral centers where the patient population may vary significantly from those studied in larger phase III trials.

Although speculative, one immediate application of synthetic datasets in phase II studies could be to generate groups of control patients that faithfully mimic the patients who are receiving active therapy in early phase clinical trials. If properly designed, these studies could be performed in a randomized, double-blind manner. Bayesian statistical methods could then be used to compare the response of patients receiving active therapy to patients enrolled in a synthetic control group. This would allow investigators to prioritize their precious resources to enroll more patients in the active therapy arms, which would also mitigate some of the statistical problems that occur when using small control groups that do not complement the demographics of the disease being studied. Another way in which synthetic data could be used is in the context of large-scale and pragmatic trials that evaluate novel targeted therapies that involve genomic targets, insofar as conventional randomized clinical trials are often impracticable because of the large sample sizes that are required to demonstrate clinical effectiveness in this setting (12,13). Lastly, one can imagine using synthetic datasets to predict trends in rare diseases, which in turn could be used to design appropriately powered clinical trials that target clinically meaningful end points.

What are some of the limitations of using synthetic data to evaluate drug responsiveness? One potentially important limitation is that whereas synthetic models derived from existing datasets may replicate certain general trends of the dataset, they may not necessarily be able to predict specific trends within a dataset (e.g., all-cause death vs. cardiovascular death). Although this limitation remains theoretical at present, it may be problematic with respect to using synthetic datasets to evaluate novel therapeutics. Whether creating a larger derivative dataset that contains an adequate number of outcomes of interest in order to estimate drug effects accurately will satisfactorily address this issue remains an important question that will require further study. Second, there is no consensus about how best to create synthetic datasets. Fully synthetic datasets do not contain any original data, whereas partially-synthetic datasets may only de-identify or anonymize sensitive values. There are theoretical advantages and disadvantages to both approaches; however, there is no information with respect to which approach is better for predicting drug responsiveness. Lastly, at the time of this writing, the Food and Drug Administration has not yet approved the use of synthetic datasets for registration studies: it is simply too soon.

Reducing the cost of developing new cardiovascular therapies will require fundamental changes to the way in which we conduct preclinical and clinical trials in order to make them faster, cheaper, and more adaptable. Here, we suggest that the use of synthetic data derivatives may help with the development of new and novel cardiovascular drugs. As always, we welcome comments and suggestions from

investigators in academia and industry, patients, societies, and all of the governmental regulatory agencies about your thoughts about the potential role of synthetic data in translational medicine, either through social media (*#JACC:BTS*) or by e-mail (JACC@acc.org).

ADDRESS FOR CORRESPONDENCE: Dr. Randi Foraker, Institute for Informatics, Washington University School of Medicine, 4444 Forest Park Avenue, Suite 6318, St. Louis, Missouri 63110. E-mail: randi.foraker@wustl.edu.

REFERENCES

1. Mann DL. The rising cost of developing cardiovascular therapies and reproducibility in translational research: do not blame it (all) on the bench. J Am Coll Cardiol Basic Transl Science 2017;2:627–8.

2. Embi PJ, Kaufman SE, Payne PR. Biomedical informatics and outcomes research: enabling knowledge-driven health care. Circulation 2009; 120:2393–9.

3. Davis BR, Cutler JA, Gordon DJ, et al. Rationale and design for the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). ALLHAT Research Group. Am J Hypertens 1996;9:342–60.

4. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003;290:1624–32.

5. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues Clin Neurosci 2011;13:217–24.

6. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. Health Aff (Millwood) 2014; 33:1229–35.

7. Mulder R, Singh AB, Hamilton A, et al. The limitations of using randomised controlled trials as a basis for developing treatment guidelines. Evid Based Ment Health 2018;21: 4–6.

8. Krumholz HM. Open science and data sharing in clinical research: basing informed decisions on the totality of the evidence. Circ Cardiovasc Qual Outcomes 2012;5:141–2.

9. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14:550–63.

10. U.S. Department of Health and Human Services. Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance With the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Washington, DC: U.S. Department of Health and Human Services, 2012 . Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed September 26, 2018.

11. Lane RS. An interview with Robert S. Lane, Ph. D. Interviewed by Vicki Glaser. Vector Borne Zoonotic Dis 2010;10:211–5.

12. De Gruttola VG, Clax P, DeMets DL, et al. Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. Control Clin Trials 2001;22:485–502.

13. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med 2011;8:161–73.