Pediatric RESEARCH

COMMENT

Privacy-preserving statistical analyses in Learning Health Systems

William Gardner[1,2]

Pediatric Research (2020) 87:978–979; https://doi.org/10.1038/s41390-020-0835-4

A Learning Health System (LHS) is one in which "internal data and experience are systematically integrated with external evidence, and that knowledge is put into practice".[1] To accomplish this goal, we will need to analyze large volumes of routinely collected health data. However, creating data sets that span clinical populations poses significant problems of privacy and data governance. The article by Toh et al.[2] demonstrates a possible way around these privacy and governance challenges.

To advance personalized medicine, we need to develop tools that can predict how the outcomes of diseases or treatments will vary based on a profile of individual patient characteristics.[3] Developing predictive models with sufficient precision to guide the tailoring of treatments to individuals requires large data sets. However, assembling large data sets is a challenge, in part because the relevant data are often held by independent stakeholders, as in the case considered by Toh, where data on BMI and antibiotic exposure have been collected by PEDSnet,[4] a data-sharing consortium of pediatric hospitals.

One way to do this is to export the data tables from each hospital and pool them in a common table (see Fig. 1, panel a). However, pooling individual data across hospital boundaries requires the fortification of the data pool to protect patient privacy, as well as procedures to control who is authorized to view the data. This is expensive and risky.

But as Toh et al. demonstrate, for analyses based on ordinary least squares regression and some generalized linear models (hereafter, "standard regression"), it is possible to analyze a multi-institution data set without pooling the data across institutions. They do this by exploiting a fact in mathematics: standard regressions do not require the analysis of individual data. You can estimate standard regressions from summary statistics (e.g., for ordinary least squares regression, the variable means and the covariance matrix). Figure 1 (panel b) illustrates this. Each hospital calculates the statistics summarizing its local data. The summary statistics are then exported and used to calculate pooled summary statistics, from which the analysts estimate the regression. Toh et al. showed that the results of the pooled individual data (panel a) and pooled summary data (panel b) approaches were identical. Although this was never in doubt, the demonstration illustrates the value of the method.

The pooled individual data analysis versus pooled summary statistics analysis contrast is closely related to the difference between individual participant data meta-analysis (IPDMA)[5] and standard meta-analysis. IPDMA pools individual-level data from all the controlled trials of an intervention to estimate a common treatment effect, while standard meta-analysis harvests means and standard deviations from each trial to the same end. Given that pooling individual participant data is expensive and time consuming, why would we ever do it? Is there ever a need to construct pooled, cross-hospital individual-level pediatric data sets?

Unfortunately, unlike standard regressions, many analyses require more than pooled summary statistics. As Toh et al. note, these analytical computations use iterative optimization algorithms that repeatedly use individual-level data. Examples include nonlinear models, models involving clustering and nesting of subjects, Bayesian statistics, and nearly every species of machine learning. Iterative optimization is often required in predictive analytics, genomics, health geography, psychometrics, and population health.

Unlike standard regressions, in these analyses you cannot estimate the parameters only from the summary statistics. Instead, you estimate them with an algorithm like this:

1. Set some start values for the parameters of your model.
2. Measure the goodness-of-fit between your current model estimates and the individual-level data.
3. Check how much the current goodness-of-fit has improved compared to the last time you tried.
4. If the improvement in the goodness-of-fit is minimal, your current parameter estimates are the solution, because they are likely as good as they will get. You can stop calculating.
5. Otherwise, you can analyze the discrepancy between the model and the individual data to calculate new parameter estimates that will fit the data better.
6. Go back to Step 2.

As can be seen, iterative optimization requires repeated evaluations of the fit between the model and the individual data. This is straightforward using pooled individual data (panel a), but it can't be done readily using the pooled summary statistic approach (panel b).

It may be possible, however, to extend Toh et al.'s approach and develop iterative algorithms that keep individual-level data protected in local hospital databases. Modern iterative optimization algorithms work in parallel: the data are partitioned and distributed across many servers, and so are large portions of the computations on those data.[6] This suggests that iterative optimization algorithms could be redesigned so that each iteration implements a version of Toh et al.'s summary statistic algorithm. The portions of the calculations that require individual data—step 2 above—could be carried out in a distributed manner within the local hospital computing environments. Then the information about the discrepancies between the model and the individual

[1]University of Ottawa, Ottawa, ON, Canada and [2]CHEO Research Institute, Ottawa, ON, Canada
Correspondence: William Gardner (wgardner@cheo.on.ca)

Received: 6 November 2019 Accepted: 11 November 2019
Published online: 14 March 2020

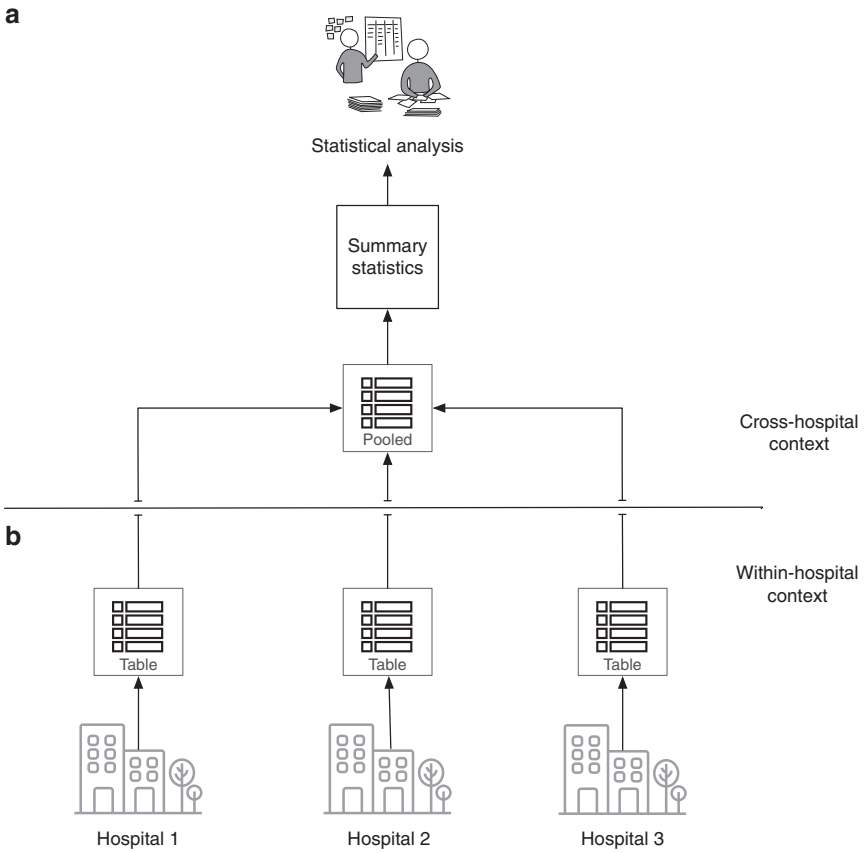

Fig. 1 Pooling Multiple Data Sets into a Common External Database. The most frequent method for building a data commons is to have multiple institutions feed their data sets into a common external store. Researchers then access the data from the external store.

data, which has no signature of the individuals, could be exported and pooled to evaluate the goodness-of-fit and improve the parameter estimates (steps 3−5). The extended algorithm would inherit the central virtue of Toh et al.'s pooled summary statistic algorithm, in that individual data would never cross the boundaries of the local hospital computing environments.

To carry out distributed iterative optimizations, the consortium of hospitals would need to implement a common informatics architecture that would allow the algorithm to transfer interim results back and forth across the boundaries of the local hospital systems. Implementing the computational architecture required to support this algorithm would be a significant commitment, but the development of information architectures to support collaborative computing has long been a goal of Learning Health Systems,[6] including the PedsNET initiative.[7]

The multi-institutional analyses demonstrated by Toh et al. are critical to the future of precision medicine and population health. The most difficult challenges are likely organizational.[8] Great efforts will be needed to get stakeholder institutions to implement common terminologies for medical data and interfaces for distributed computation, and to sustain them. The PedsNET collaborators are pioneers in these efforts.

ACKNOWLEDGEMENTS

The author's research is supported by grants from the Canadian Institutes of Health Research and the ScotiaBank Foundation, and by research contracts with the Public Health Agency of Canada and the US Centers for Disease Control.

AUTHOR CONTRIBUTIONS

The author is solely responsible for this text.

ADDITIONAL INFORMATION

Competing interests: The author declares no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Agency for Healthcare Research and Quality. *About Learning Health Systems [Internet]* (Agency for Healthcare Research and Quality, Rockville, MD, 2019).
2. Toh, S. et al. Privacy-protecting multivariable-adjusted distributed regression analysis for multi-center pediatric study. *Pediatr. Res.* (2019). https://doi.org/10.1038/s41390-019-0596-0. [Epub ahead of print].
3. Fröhlich, H. et al. From hype to reality: data science enabling personalized medicine. *BMC Med.* **16**, 150 (2018).
4. Forrest, C. B., Margolis, P., Seid, M. & Colletti, R. B. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. *Health Aff.* **33**, 1171–1177 (2014).
5. Stewart, L. A. & Clarke, M. J. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Stat. Med.* **14**, 2057–2079 (1995).
6. Mandl, K. D. et al. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. *J. Am. Med. Inf. Assoc.* **21**, 615–620 (2014).
7. Forrest, C. B. et al. PEDSnet: a National Pediatric Learning Health System. *J. Am. Med. Inf. Assoc.* **21**, 602–606 (2014).
8. Mandl, K. D. & Kohane, I. S. Federalist principles for healthcare data networks. *Nat. Biotech.* **33**, 360–363 (2015).