***Editors' note:** This series addresses topics of interest to epidemiologists across a range of specialties. Commentaries start as invited talks at symposia organized by the Editors. This paper was presented at the 45th Annual meeting of the Society of Epidemiologic Research (SER) in Minneapolis, MN, 2012.*

Is Size the Next Big Thing in Epidemiology?

Sengwee Toh and Richard Platt

A recent assessment of drugs that target the renin-angiotensin-aldosterone system and angioedema risk drew from a source population of more than 100 million people and 350 million person-years of observation time. The assessment identified 3.9 million eligible new users of angiotensin-converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), the direct renin inhibitor aliskiren, or the common referent group beta-blockers (a class of drugs not thought to affect the risk of angioedema). More than 4500 outcome events were observed.[1] The assessment replicated a well-known association between ACEIs and angioedema,[2–4] but the risk estimates were much more precise than those from prior studies. The assessment also generated new evidence for ARBs and aliskiren.

Not so long ago, an assessment of such scale existed only in our imaginations. Secondary uses of routinely collected electronic health information now enable us to conduct research using data from hundreds of thousands or even millions of patients.[5] But certain studies or surveillance activities, especially those with rare exposure or outcome, demand data larger than any single extant source. Combining data from multiple sources would help solve the sample size problem, but sharing data has always been a challenge because of privacy, security, regulatory, legal, and proprietary concerns. How did the angioedema assessment accomplish this and what implications does it have for epidemiology?

DISTRIBUTED DATA NETWORKS TO SUPPORT LARGE-SCALE EPIDEMIOLOGIC ASSESSMENTS

The angioedema assessment was conducted within a distributed network of electronic healthcare databases created as part of the Mini-Sentinel program, a pilot project funded by the US Food and Drug Administration to conduct postmarket surveillance of medical product safety.[6,7] Distributed data networks are not new. The HMO Research Network and its research consortia have been in existence for years.[8–12] Other examples include the Observational Medical Outcomes Partnership[13] and the EU-ADR project.[14] The Office of the National Coordinator for Health Information Technology's Query Health program promotes the use of distributed systems for secondary uses of electronic health data.[15]

Compared with the alternative of storing all data in a centralized location, a distributed data system is more appealing to most data holders because they maintain direct control over their data.[16–18] In a typical epidemiologic study within a distributed data network, the lead team develops and tests the analytic program—often against a common data model (more below)—and then distributes it to all participating sites. The type of outputs returned to the requester varies by study, ranging from simple counts to an individual-level analytic dataset.

From the Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA.
Supported, in part, by the US Food and Drug Administration through the Department of Health and Human Services contract no. HHSF223200910006.
Correspondence: Sengwee Darren Toh, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 133 Brookline Ave 6th Floor, Boston, MA 02215. E-mail: darrentoh@post.harvard.edu.

ISSN: 1044-3983/13/2403-0349
DOI:10.1097/EDE.0b013e31828ac65e

An obvious advantage of having a large sample size from multiple data sources is the ability to study rare exposures, outcomes, or subgroups. For example, because aliskiren was approved relatively recently, in 2007, its analysis in the angioedema assessment would not have been possible without a large source population. Data from multiple sources also often represent demographically and geographically diverse study populations, which allow for study of geographic or practice variations or treatment heterogeneity. However, these advantages incur methodological and operational complexity, as we describe below.

Methodological Considerations

There is often a trade-off between analytic flexibility and the granularity of information that needs to be shared. Highly summarized data preserve privacy and confidentiality, but limit what we can do analytically. Sharing detailed individual-level data allows for more sophisticated analysis—and in most cases, effectively turns the study into a single-dataset study—but may raise concerns over privacy, security, and proprietary information. Balancing these concerns against analytic rigor can be difficult.

Recent methodological advances have given us new analytic tools to perform complex statistical analysis without the need to share potentially identifiable information.[19,20] The angioedema assessment used propensity scores to adjust for confounders. Propensity scores,[21] like another commonly used confounder summary score—disease risk scores,[22] have properties that are particularly useful for distributed analyses: they can be computed locally by the data partner, and they condense information on a large number of confounders into a single, nonidentifiable measure, thus eliminating the need to transfer highly granular, potentially identifiable individual-level information. The angioedema assessment further used two analytic methods to incorporate propensity scores into the analysis that avoided the sharing of any individual-level data: a case-centered logistic regression approach (which requires only risk-set data)[23] and an inverse variance-weighted meta-analysis (which requires only site-specific effect estimates and standard errors).[20] Mathematical proof and empirical studies have shown that these two methods provide results that are identical or similar to the results obtained from individual-level data analysis.[20,23]

Additional methods that do not require sharing of potentially identifiable information include distributed regression[24] and matching or stratification by confounder summary scores.[19] Distributed regression fits regression models on distributed databases and produces results identical to those from analysis of individual-level data. Sites transfer only summary statistics for model fitting. In some studies, it may be sufficient to perform all analyses by stratifying or matching—within data source—on a confounder summary score and sharing the aggregated stratum-specific or match-set data.

To perform confounder adjustment, with or without confounder summary scores, we can have each site adjust for the same covariates. The advantage of this "common denominator" approach is consistency, but it does not fully use additional confounder information available only at certain sites. An alternative would be to have each site adjust for its own set of covariates. This approach reduces residual confounding but is operationally more cumbersome because there can be no "one-size-fit-all" distributed analytic program. A special case of the site-specific approach is the high-dimensional propensity score method, which allows for prespecification of a common set of covariates and empirical identification of additional site-specific covariates.[25,26] It is possible to develop a distributed program to perform such an analysis.[27] In a given study, multiple approaches can be used to examine the robustness of findings.

Data Considerations

In our experience, it is most efficient for all data partners in a distributed network of electronic healthcare databases to extract, transform, and load their source data into a standard format that conformed to a common data model before running any analysis. A common data model ensures uniform data-file structures and data-element naming conventions and definitions.[16–18] It allows data checking, manipulation, and analysis via identical computer programs developed and tested by few but shared by all, reducing programming burden and the chance for errors across sites. In the angioedema assessment, for example, all Mini-Sentinel data partners had transformed their source data using a common data model[28] and passed data-quality checks before the assessment. The angioedema project team developed, tested, and distributed analytic programs to all data partners, who ran the programs and returned aggregate data to the team to complete the analysis.

It is crucial to engage data partners at all steps of any studies. When creating a common data model and transforming the source data, local data expertise is needed to identify issues related to coding practice, institutional practice guideline, clinical workflow, and other factors that may affect the completeness and accuracy of data. Some data issues are systematic and relevant for all data partners, whereas others are idiosyncratic and site-specific. Data partners play a central role in checking the quality of the data, both when data are added to the master distributed dataset and during project-specific analyses. They are also involved in obtaining additional material (such as medical records) to confirm information in the electronic data when necessary.[29] The Mini-Sentinel common data model would not exist without contributions from local data experts and investigators with extensive experience analyzing their data.

Where Do We Go from Here?

We need to learn more about the strengths and limitations of existing methods in various study settings. We also need to improve these methods and develop new ones that are more analytically robust and operationally efficient in distributed data networks. For example, the distributed regression approach currently allows linear and logistic regression, but it needs to be extended to include other regression models. Methods that can incorporate confounder summary scores seem promising, but we need to better understand—in the context of distributed environments—the advantages and disadvantages of using these scores compared with treating each confounder individually, when to select between propensity scores and disease risk scores, and how to best estimate these scores and perform model diagnostics.

The angioedema assessment joins a growing body of real-world experience that demonstrates how distributed networks of electronic healthcare databases support large-scale epidemiologic studies or surveillance assessments. So, is size the next big thing in epidemiology? Our answer is, "Sometimes." A large sample size does not solve every epidemiologic problem. For one, it does not make unmeasured confounding go away. Sometimes it is more prudent to get more information about a smaller group of patients than to study an additional million. But having a large sample size allows us to do a lot more, and sometimes it is just what we need.

ACKNOWLEDGMENTS

We thank Sean Hennessy of University of Pennsylvania and Carlos Bell, Melissa Robb, and Marsha Reichman of the US Food and Drug Administration for their thoughtful comments on an earlier draft of this commentary.

REFERENCES

1. Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. *Arch Intern Med*. 2012;172:1582–1589.
2. Vleeming W, van Amsterdam JG, Stricker BH, de Wildt DJ. ACE inhibitor-induced angioedema. Incidence, prevention and management. *Drug Saf.* 1998;18:171–188.
3. Brown NJ, Ray WA, Snowden M, Griffin MR. Black Americans have an increased rate of angiotensin converting enzyme inhibitor-associated angioedema. *Clin Pharmacol Ther*. 1996;60:8–13.
4. Miller DR, Oliveria SA, Berlowitz DR, Fincke BG, Stang P, Lillienfeld DE. Angioedema incidence in US veterans initiating angiotensin-converting enzyme inhibitors. *Hypertension*. 2008;51:1624–1630.
5. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58:323–337.
6. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System—a national resource for evidence development. *N Engl J Med*. 2011;364:498–499.
7. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):1–8.
8. Andrade SE, Raebel MA, Boudreau D, et al. Chapter 12: Health maintenance organizations/health plans. In: Strom BL, Kimmel SE, Hennessy S, eds. *Pharmacoepidemiology*. 5th ed. Chichester, UK: Wiley-Blackwell; 2012.
9. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiol Drug Saf.* 2001;10:373–377.
10. Wagner EH, Greene SM, Hart G, et al. Building a research consortium of large health systems: the Cancer Research Network. *J Natl Cancer Inst Monogr*. 2005; 3–11.
11. Go AS, Magid DJ, Wells B, et al. The Cardiovascular Research Network: a new paradigm for cardiovascular quality and outcomes research. *Circ Cardiovasc Qual Outcomes*. 2008;1:138–147.
12. Chen RT, Glasser JW, Rhodes PH, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. *Pediatrics*. 1997;99:765–773.
13. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010;153:600–606.
14. Coloma PM, Schuemie MJ, Trifirò G, et al; EU-ADR Consortium. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf.* 2011;20:1–11.
15. Office of the National Coordinator for Health Information Technology. Query Health Project Charter. Available at: http://QueryHealth.org. Accessed 19 September 2011.
16. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. *Ann Intern Med*. 2009;151:341–344.
17. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care*. 2010;48(6 suppl):S45–S51.
18. Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. *Clin Pharmacol Ther*. 2011;90:883–887.
19. Rassen JA, Moran J, Toh D, et al. Evaluating strategies for data sharing and analyses in distributed data setting. http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_Evaluating-Strategies-for-Data-Sharing-and-Analyses.pdf. Accessed 28 February 2013.
20. Rassen JA, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care*. 2010;48(6 suppl):S83–S89.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
22. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res*. 2009;18:67–80.
23. Fireman B, Lee J, Lewis N, Bembom O, van der Laan M, Baxter R. Influenza vaccination and mortality: differentiating vaccine effects from bias. *Am J Epidemiol*. 2009;170:650–656.
24. Karr AF, Lin X, Sanil AP, Reiter JP. Secure regression on distributed databases. *J Comput Graph Stat.* 2005; 14:263–279.
25. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20:512–522.
26. Toh S, García Rodríguez LA, Hernán MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiol Drug Saf.* 2011;20:849–857.
27. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiol Drug Saf.* 2012;21(suppl 1):41–49.
28. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf.* 2012;21(suppl 1):23–31.
29. Cutrona SL, Toh S, Iyer A, et al. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. *Pharmacoepidemiol Drug Saf.* 2013;22:40–54.