Editor’s Perspective

Of the Importance of Motherhood and Apple Pie
What Big Data Can Learn From Small Data

Véronique L. Roger, MD, MPH

Data integrity and validity are challenging topics to get published in medical journals. The clinical implications of these matters are not always intuitive to the reader, and data quality seldom makes headlines in the medical literature, unless something goes wrong, in which case-related conversations can make their way to the court rooms and the lay press. Yet, the cornerstone of our ability to make robust inference and sound clinical decisions is the assumption of the validity, accuracy, and representativeness of medical research data. Over the past few years, several articles in *Circulation: Cardiovascular Quality and Outcomes* have reminded us of the importance to formally examine such assumptions.

Registries occupy a central role in the clinical research landscape by allowing to pool observational clinical data into larger data sets to enhance statistical power and to draw more robust inference. Registries afford expansive population coverage and contain abundant clinical data that enable valuable risk adjustment. Most importantly, registries are intended to reflect real-world practice. However, as participation is voluntary, the representativeness of patients in registries cannot be assumed as institutions that participate may differ from those that elect not to. Furthermore, the completeness of the data must be verified as illustrated by a study of the Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes with Early Implementation of the ACC/AHA Guidelines (CRUSADE) National Quality Improvement Initiative pertaining to patients with non-ST-segment-elevation acute coronary syndromes, where it was identified that records often lacked key clinical elements of the history and physical examination. Hence, the quality control of registries is critical to their value for outcomes research and clinical care. Regular audits of registries to assess their quality are thus essential to their use.

In *Circulation: Cardiovascular Quality and Outcomes*, Ferreira et al reported on an audit of a national acute coronary syndrome registry. Patients who were included in the registry were less likely to die in hospital and more likely to receive recommended treatments than those who were not included. Furthermore, among 129 registries on acute coronary syndromes, audits were seldom mentioned for recruitment/inclusion (29%) or data abstraction (37%). More recently, also in *Circulation: Cardiovascular Quality and Outcomes*, Reeves et al evaluated the accuracy of the Michigan Stroke Registry by linking to a state-wide hospital discharge database (considered as the gold standard). The Michigan Stroke Registry is a part of the Paul Coverdell Acute Stroke Registry, a multistate hospital-based registry and quality improvement program overseen by the Centers for Disease Control and Prevention. Among 26 hospitals in the Michigan Stroke Registry, the sensitivity of case ascertainment was 69% and the positive predictive value was 89%. The sensitivity was lower in teaching hospitals and primary stroke centers, but was higher in the sites that used prospective ascertainment. Through this process, some hospitals that had changed their ascertainment method were identified, illustrating that, much like laboratory assays, registries can drift as several hospitals had changed their procedures in a way that substantially impacted the integrity of the data without knowledge of the registry staff. Overall, an evaluation of 153 national registries in the United States delineated substantial opportunities to enhance the robustness of registries and the quality of data collection and analyses to optimize their clinical use. Strikingly, only 18% of registries indicated that they audit the data. These articles provide a valuable illustration of the importance of methodological discipline applied to registries: it is our responsibility as researchers and as clinicians to make sure that the data we collect are accurate and complete and that it remains that way over time. Registries must implement and report regular quality control measures to optimize the quality of their data, which must be evaluated by conducting and periodically repeating audits.

Administrative data are another cornerstone of the outcomes research toolbox, which increasingly relies on administrative data using *International Classification of Diseases-Ninth Edition* codes for case finding and often ascertainment. In 2014 of *Circulation: Cardiovascular Quality and Outcomes*, Kumamaru et al reported on the validity of diagnostic coding algorithms for identifying stroke in the Medicare population by linking data from the Reasons for Geographic And Racial Differences in Stroke (REGARDS) study to Medicare claims. Expert-adjudicated strokes from medical records served as the gold standard. Claims-based algorithms had high positive predictive value and specificity to identify stroke supporting their use as outcomes. However, these algorithms were unsuitable to estimate stroke incidence because of low sensitivity. More recently, also on the same topic and in *Circulation: Cardiovascular Quality and Outcomes*, Thigpen et al reported on the accuracy of *International Classification of Diseases-Ninth Edition* codes to identify stroke, and atrial fibrillation compared with medical record review. Reliance on
International Classification of Diseases-Ninth Edition codes to identify patients with stroke and atrial fibrillation resulted in notable inaccuracies of similar magnitude across 3 different institutions: Boston Medical Center, Geisinger Health System, and the University of Alabama. Taken collectively, these studies underscore that administrative data without event validation should be interpreted with caution and that the appropriateness of their use will depend on the goal of the study.

The critical principles of data accuracy and integrity are conceptually undisputable and universally agreed on, much like the proverbial motherhood and apple pie. Why is it important to reflect on these principles at this time? First, because as any accepted rule, they can be taken for granted and forgotten. The aforementioned examples clearly illustrate that the value of registries to advance science and improve care is at stake if rigorous quality control is not implemented. As the appetite for registries grows both for discovery research and for quality improvement, standard operating procedures must ensure that these critical quality control steps are regularly applied. Centers for Medicare and Medicaid Services created a program entitled, Qualified Clinical Data Registries, which compiles registries in which eligible practitioners can participate. This is a laudable effort to optimize the quality of registries that should pave the road for widespread efforts, such as the steps deployed in the National Cardiovascular Data Registry.

Second and more importantly, these timeless principles have to be explicitly brought to the forefront of our collective thinking as clinical and population sciences research enter the brave new world of Big Data. Regardless of the definitions, often disparate and vague, that surround the concept of big data, thinking as clinical and population sciences research enter the brave new world of Big Data. Regardless of the definitions, often disparate and vague, that surround the concept of big data, the potential of new tools to collect and analyze larger data sets is intuitively recognized by most, and large-scale repositories, disease-agnostic or disease-specific, are proliferating, and the appeal of electronic data is amplified by the current fiscal constraints. On the healthcare delivery side, in recent years, providers have been incentivized to shift from paper to electronic medical records (EMR), thereby greatly increasing the volume and accessibility of health data available to clinicians, researchers, and patients. Furthermore, other data sources, which encompass mobile health devices, claims, and administrative databases, expand the scope of the data that can be collected and linked to EMRs. These are unprecedented opportunities to not only facilitate the capture of exposure and outcomes in a conventional hypothesis-driven approach but also mine these vast data repositories agnostic to any hypothesis after an in silico discovery science approach applied to large volumes of clinical and health data. These are incredibly thrilling promises, for some akin to moonshots.

To realize this exciting potential, the optimal methods to acquire, validate, standardize, and analyze high volumes of data generated by different sources must be delineated, codified, and broadly shared with the scientific community.

Recommendations to move forward in this direction have been put forth by several organizations including the National Cancer Institute and the National Heart Lung and Blood Institute. These recommendations vigorously underscored that determining the validity/accuracy/integrity of data is critical steps to progress from collecting data to generating information and deriving knowledge. Much remains to be done to execute on this directive. As far as EMRs, cautionary reports on their use for research and performance/quality assessment are surfacing. As we acknowledge that more work is needed to improve our operational understanding of the research applications of EMRs, we nevertheless conceptually understand the limitations of the use of medical records for research, a topic familiar to most providers and clinical researchers. Indeed, the migration to electronic platforms cannot be expected to alleviate issues related to missing data, uneven reliability, incomplete documentation, and inconsistencies across providers and over time. Analogous limitations pertaining to completeness of records have also been reported to some extent for claims data but could benefit from more attention to fully delineate the best methods to be applied to and the most appropriate research questions to be addressed by claims data.

The complexity of the questions increases significantly when one considers emerging data sources. What does data integrity, validity, and representativeness mean for person-generated health data, including from mobile devices, social media platforms, and other patient-generated data sources? In this regard, the early experience of the Beacon Community Cooperative Agreement Program created by the Office of the National Coordinator for Health Information Technology is particularly informative. Several Beacon Communities launched community-based mobile health programs and shared useful lessons learned. These early considerations, including the recognition of key barriers to mobile health use, some related to socioeconomic status, underscore the pressing need to address among other topics, the implications of the digital divide on health disparities.

Although many questions remain unanswered and much work remains to be carried out, it is imperative to do so. Bad data are costly to business as Experian underscored, suggesting that 75% of businesses are wasting and 14% of revenue is because of poor data quality. If this is the case for business, one can only imagine the magnitude of the implications for clinical research and health care. Privacy issues add to the complexity of the discussion. Indeed, responsible data sharing is essential to optimize the quality of the information on the effectiveness and safety of medical care. To delineate the good common practices for responsible data sharing, the perspectives of key stakeholders must be considered. For research participants, informed consent, privacy protection, and knowledge to be gained from data sharing are essential. Researchers seek appropriate time to report their findings and hence get academic credit for their work. Considerations related to intellectual property and commercialization also come into play. Forging agreements on measures to address these concerns from these diverse perspectives will be difficult but is vital.

In conclusion, it is obvious from this brief perspective that many thoughtful conversations and formal research studies are urgently needed on these topics. These endeavors must unfold in settings that foster broad engagement of key stakeholders including clinicians, clinical researchers, methodologists, patients, and their families. Finally, robust action plans must follow to be deployed with alacrity. At Circulation: Cardiovascular Quality and Outcomes, we look forward to taking an active part in this exciting journey.
Disclosures

None.

References


Key Words: data science • epidemiology • population • registries
Of the Importance of Motherhood and Apple Pie: What Big Data Can Learn From Small Data

Véronique L. Roger

*Circ Cardiovasc Qual Outcomes*. 2015;8:329-331; originally published online July 14, 2015; doi: 10.1161/CIRCOUTCOMES.115.002115

*Circulation: Cardiovascular Quality and Outcomes* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Copyright © 2015 American Heart Association, Inc. All rights reserved.

Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:

http://circoutcomes.ahajournals.org/content/8/4/329

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Circulation: Cardiovascular Quality and Outcomes* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at: http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Circulation: Cardiovascular Quality and Outcomes* is online at: http://circoutcomes.ahajournals.org//subscriptions/