Infrastructure Requirements for Secondary Data Sources in Comparative Effectiveness Research

Demissie Alemayehu, PhD, and Jack Mardekian, PhD

The growing interest in comparative effectiveness research (CER) has re-ignited the debate about the inadequacy of data from randomized controlled trials (RCTs) to address patient-centered decision making. Despite their well-known internal validity and use as the gold standard for regulatory decision making, the limitations of RCTs are widely recognized. In addition to their lack of statistical power due to inadequate sample size to address certain research hypotheses, practical and ethical considerations may preclude their viability. A case in point is the ethical dilemma in conducting an RCT to establish whether a diet high in fat content may be a risk factor for dementia, which might produce useful public health information but would not be acceptable in terms of protection of human subjects. Frequently, RCTs provide substantial information regarding the efficacy of drugs and other medical interventions, yet leave large gaps in evidence that would be relevant for medical decision making. Even when RCTs are carried out with this intent, they may not necessarily reflect “real-world” experience and, therefore, may not provide sufficient evidence to guide patient-centered care.

The substantial investment in CER and the broad objective implied in the American Recovery and Reinvestment Act of 2009 have necessitated the need to seek alternative sources of data to meet the emerging health care questions. The stated requirements include an “…assessment of a comprehensive array of health-related outcomes for diverse patient populations and sub-groups” as well as “…a wide range of interventions;” and “[development, expansion, and use of a variety of data sources and methods to assess comparative effectiveness.”1-2

Contending with the changes in health care policy and delivery clearly requires doing things differently, as discussed at a recent workshop sponsored by the Institute of Medicine.3 The workshop summary highlighted the dependence on clinical trials as the “sole source of evidence on the constantly accelerating flow of diagnostic and treatment challenges is unfeasible.” The need for a “learning healthcare system” with “real-time learning from the clinical experience and seamless application of the lessons in the care process” was emphasized.

Secondary data, such as registries and retrospective databases, can be used to complement RCTs, since they are less costly and can be used to incorporate real world experience to answer questions. Further, important questions, such as adherence, treatment patterns, and burden of disease, can be answered in retrospective analyses of databases. However, effective use of secondary data requires addressing major methodological and infrastructural issues that may be related to, but often go beyond, those encountered with most RCT-based work. Infrastructural issues, such as tools to efficiently access and correctly analyze the data, need to be developed for effective use of such data sources. Guidelines need to be formulated and data standards established using RCTs as a role model. Data warehouses are also required to be established that respect the privacy and confidentiality of patients.

In this paper, we discuss the infrastructural requirements for secondary data utilization in CER, and identify gaps that must be filled to address the underlying issues, with emphasis on data standards, data quality assurance, data warehouses, computing environment, and protection of privacy and confidentiality.

Secondary Data Sources and Associated Challenges

Secondary data can be generated from registries, chart reviews, electronic health records (EHR), administrative claims databases, or national surveys such as the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System (BRFSS). There may be linkage between distinct sources (e.g., U.S. Renal Data System [USRDS]) registry of end-stage renal disease patients in which claims data from Medicare patients are linked).4-6

In other cases, the source may involve alternative designs (e.g., a combination of RCTs and nonrandomized studies). For instance, the United Kingdom’s General Practice Research Database (GPRD) has been developing the capability to run real-world primary care clinical trials through recruitment at point of care by general practitioners in their system. Patients are recruited at point of care. Software from GPRD’s information technology (IT) system informs the doctor when a patient satisfies inclusion and exclusion criteria for a particular study protocol, prompts the doctor for other needed information as well as patient consent, and provides the randomization (if there is one) to an assigned drug group. Drugs are given open label. Patient follow-up can be according to the treating physician’s standard care or can be a specific prompted return to the doctor. All data, including safety and patient reported outcomes, are recorded in the standard EHR data downloaded on a regular basis by GPRD.

Secondary data sources offer several benefits. First, they are rooted in real life, and if analyzed using appropriate statistical methods can document the effectiveness of drugs in everyday use under a wide spectrum of clinical practice. The patient population is diverse, mimicking the real world. Further, this patient diversity allows the rapid identification of large numbers of patients in a cost-effective manner, and permits comprehensive, long-term safety follow-ups. The latter is particularly important when the focus is on rare diseases, atypical therapy responses, or uncommon clinical questions.
However, observational studies also have inherent methodological and infrastructural limitations. Statistical methods have been developed to mitigate the limitations, and guidance documents have been generated to upgrade the analysis and reporting of data from secondary sources.

From the operational perspective, the infrastructural limitations of secondary data are considerable. In general, all relevant data may not be available. For example, reasons for therapeutic substitution may not be known, actual low-density lipoprotein cholesterol (LDL-C) levels for statin users may not be tracked, or diagnoses associated with a medication prescription may not be recorded. Further, claims data are generally built for billing and record-keeping purposes, and not for research. Therefore, the potential for error occurs at many points along the record keeping process. The implication for researchers is that both systematic and random error can occur in the identification of treatment exposure and outcome. In addition, there is currently no simple approach to link the health information of patients from separate data sources. A case in point is the inability of insurers to readily link laboratory results with patient information from separate pharmacy plans. In addition to logistical constraints, the risk for re-identification of patients increases as the amount of information increases.

**Infrastructural Requirements**

3.1 Data Standards. Lack of standardized data limits the analyst's ability to efficiently process data, implement standard statistical packages, integrate analysis results, and report results with transparency. Progress in defining standards in secondary data is slow relative to the progress made in defining standards in the clinical trial world, which is largely due to the efforts of the Clinical Data Interchange Standards Consortium (CDISC). CDISC, founded almost 10 years ago, is "a global, open, multidisciplinary, nonprofit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata." Standards established by CDISC are intentionally "vendor-neutral, platform-independent, and freely available," and seek to optimize workflow from protocol authoring to final study reports and regulatory submission.

The extension of CDISC standards to secondary data is reasonable. CDISC's Healthcare Link project, which started in 2005, is an initiative that specifically focuses on the mission of "interoperability between health care and clinical research." This effort has established the capability "to collect relevant data from the EHR for critical secondary uses such as safety reporting (and bio-surveillance), clinical research, and disease registries."

One example of standardization in secondary data used by major U.S. providers of administrative claims databases is the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), which is designed to code and classify diagnoses from inpatient and outpatient records. Prescription drugs and insulin products are coded using the National Drug Code (NDC) scheme that is maintained by the U.S. Food and Drug Administration (FDA).

Although the coding of diseases and drugs is highly standardized, more effort is needed in defining diseases through the specification of codes. Analyses of claims databases vary in their application of coding. For example, a patient with fibromyalgia may be identified as having either 1 or 2 medical claims with diagnosis code ICD-9-CM 729.1. Another definition may include a requirement that in addition to ICD-9-CM 729.1, the patient has filled at least 1 prescription for a drug indicated for fibromyalgia during a defined time period. In CER, standardized definitions of common diseases and conditions enable the comparison of results across studies.

The announcement by Google to retire Google Health in 2012 underscores the fact that creating a standardized infrastructure for needed health information is not an easy problem to solve. Google established Google Health in 2006 as a personal health information centralization service. The service allowed Google users to merge potentially separate health records into 1 centralized profile either manually or through partnered health services. Lack of widespread adoption was the reason provided by Google for abandoning the project.

3.2 Computing Environment. A reliable and efficient computing infrastructure, including hardware, software, and support staff, is fundamental to the success of performing CER with secondary data sources. The computing environment must address data acquisition, storage, and integration in addition to housing analytical tools for data mining and analysis. The data need to be dynamically maintained over time with updated data and links to other data sources in the presence of increasing numbers of users. Understanding adherence and treatment patterns for patients on new drugs and treatments as they become available is an important capability. Therefore, it is important for suppliers of administrative claims databases to be able to provide adjudicated claims data in a timely fashion.

Data warehouses consisting of high-quality clinical trial data, administrative claims databases, and registries from various sponsors including industry, federal health agencies, and health care providers are possible and need to be established, maintained, and updated easily with new studies in a timely manner. The data do not necessarily need to be aggregated in a single warehouse and can remain in their existing secure environments using recent advances in database structure and high-speed computing to link across data sources. The U.S. Department of Health and Human Services (DHHS) is creating a “multi-payer claims database” that would combine claims data into a distributed warehouse from a range of public and private payers. The FDA is creating a similar infrastructure...
for its Sentinel System, which will enable FDA to monitor the safety of drugs and other medical products with the assistance of a wide array of collaborating institutions from a range of academic medical centers, health care systems, and health insurance companies.\textsuperscript{20}

The FDA sanctioned a clinical trial data repository known as Janus to enable FDA and the pharmaceutical industry to look retrospectively at clinical trial data and also prospectively to design future clinical trials.\textsuperscript{21} Janus is a highly structured data warehouse of clinical trial data based on the CDISC Study Data Tabulation Model and is characterized by containing information on large cohorts of patients.

Administrative claims databases are highly structured data warehouses and typically contain information on large cohorts of patients followed over long periods of time. More generally, secondary data sources are characterized by large numbers of records that require extensive data processing during analyses. For example, medication records may need to be sorted by patient and by prescription fill date in order to merge with outpatient visit records that also must be sorted by patient and diagnosis date. Even in the presence of highly structured data, other aspects of the data, such as the timing of office visits to a physician, may require processing large numbers of patient records for analyses. A retrospective database study including millions of patient visits, for example, may require summarizing cardiovascular events that occur at 3 and 6 months after the start of a drug therapy. The schedule of visits according to usual care practices necessitates establishing visit windows and extensive data processing to classify the cardiovascular events into the defined windows for analysis. In contrast, an RCT protocol visit schedule is aligned with its objectives, with visits occurring at periodic intervals to enable analysis of outcomes at pre-specified time points.

A data warehouse of secondary data sources needs to be accessible by all users, many of whom are performing intensive computations at the same time. Users should be able to generate extracts containing different types of secondary data, such as claims data or EHR or both, for further analyses both quickly and easily rather than having to rely on a small group that has extract responsibility. Adequate disk space can be an issue. A typical extract for a retrospective database study that involves 100,000 patients and their pharmacy claims, medical procedures, and clinical diagnoses that is generated by 1 analyst might be as large as 30 gigabytes.

Cloud computing has become a viable option for secondary data sources in health care in the past few years. The National Institute of Standards and Technology (NIST)\textsuperscript{22} defines cloud computing as a “model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”

One example company developing this technology to provide a secure cloud-computing platform that is specialized for the health care industry is Explorys.\textsuperscript{23} This technology’s target is to perform queries on a data repository that consists of over 10 million patients and billions of clinical events in a Google-like manner, efficiently and quickly.

Data warehouses demand high performance analytical tools, methods, and best practices for data visualization, simulations, and analyzing data. Analytical tools should be flexible to be used with various secondary data sources with little or no modification. A software tool for querying data from a database vendor should be able to be used on data obtained from a payer without major effort.

An example of the direction of analytical software development for CER toward user access to data to generate queries and perform basic analyses is the selection of Thomson Reuters by the U.S. DHHS. The company was selected “to develop a secure, interactive tool that will enable researchers to perform comparative effectiveness studies without the need for professional computer programming.”\textsuperscript{24} While completing the project, Thomson Reuters will develop a pilot system linking multiple health care data sources. The company will test the pilot system by conducting 2 high-priority analyses on care delivery options for selected medical conditions.

Medical dictionary diagnosis and procedure coding browsers, such as EncoderPro\textsuperscript{25} from Ingenix and drug product browsers, should be used to establish common definitions of diseases and outcomes through diagnoses, procedures, and drugs. It is not uncommon for RCTs to have centralized adverse event coding so that investigator terminology is coded consistently from study to study, which is especially important for regulatory submissions. Tools to track projects can help so that similar research questions can be answered efficiently and in a consistent manner.

One example of analytical tool development is software evolved from the application of Classification and Regression Trees. Secondary data sources are useful to identify individualized patient subgroups where outcomes are optimal. Analytical software that identifies these subgroups has been developed and relies not only on access but also computing power.\textsuperscript{26}

### 3.3 Good Practices and Quality Assurance

It is important to establish internal and external processes to ensure quality, efficiency, and transparency. Protocols for studies involving secondary data sources should be registered and study results should be posted in a similar fashion as RCTs on the U.S. National Institutes of Health registry and results database (ClinicalTrials.gov) of federally and privately supported clinical trials conducted in the United States and around the world.\textsuperscript{27} Increased transparency should reduce the potential bias of study sponsors and improve the acceptance of results from studies involving secondary data sources.
Quality guidelines exist to offer direction on good practices and assuring quality when using secondary data sources. Examples include the recommendations of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR)\textsuperscript{7-9} and the International Society of Pharmacoepidemiology,\textsuperscript{13} as well as the STrengthening the Reporting of Observational studies in Epidemiology (STROBE) statement,\textsuperscript{11} the recently published Good Research for Comparative Effectiveness (GRACE) Principles,\textsuperscript{28} and numerous other resources for evaluating nonrandomized studies of comparative effectiveness.\textsuperscript{29-31}

**Protection of Patient Privacy**

A framework to improve the infrastructure to collect and share secondary data should have a provision to address the privacy concerns of patients in a transparent and credible way, and in accordance with current applicable laws such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA, Title II).\textsuperscript{32} Ambiguity in this regard will limit the voluntary and active participation of patients, and also discourage health care providers to contribute data toward this effort. Researchers, patient advocacy groups, and legislators should work together to ensure that there is a viable consensus among the various stakeholders. The need for researchers to get access to critical data should be carefully weighed against patients' rights to privacy.

In the study of rare diseases, which tend to affect vulnerable populations, particular attention should be paid to relevant policies and requirements relating to patient privacy. When the rare disease under study is associated with special prognoses or visible phenotypes, de-identification of data alone may not be adequate to guarantee anonymity. In these circumstances, processes should be in place to prevent re-identification of patients when combining data from alternative sources.

One major reason for inadequate participation by patients and excessive concern for privacy may be lack of awareness on the part of patients and providers about the underlying purpose of the CER initiative. Therefore, it would be worthwhile to make efforts to educate the public about the scientific merit of the CER initiative, and the implications for health care utilization. In this regard, institutions, such as the Patient-Centered Outcomes Research Institute, can play a constructive role by publicizing the overarching goals of CER vis-à-vis patients' need for privacy.

**Discussion**

The lofty goal of CER to promote high-quality health care to patients can be achieved mainly through the acquisition of reliable scientific information that helps health care providers, patients, and policymakers to determine the most optimal strategy for health care delivery. This in turn is predicated on the establishment of reliable infrastructure for data access, analysis and reporting, particularly when the sources of information are nonrandomized studies.

In this paper, we considered relevant gaps that must be filled to address the issues, with particular emphasis on data standards, data quality assurance, data warehouses, software requirements, and protection of privacy and confidentiality. There are needs to develop tools to readily access and correctly analyze the data; satisfy requirements to formulate guidelines to enhance quality and transparency; establish data standards using RCTs as a role model; and create data warehouses that respect the privacy and confidentiality of patients. Further, the infrastructure should leverage cutting-edge technology and permit implementation of state-of-the-art data analytical tools.

Given the scope of the problem, strong collaboration among stakeholders is critical to address the issues effectively and efficiently. This may involve establishment of processes to link and share databases, and the harmonization of hardware and software to facilitate the exchange of information among various health care entities. The collaboration may also need to involve the creation of a framework to overcome logistical impediments as well as proprietary constraints to access of information for effective systematic reviews and analyses of randomized and nonrandomized studies involving RCTs and secondary data sources. In this respect, the Observational Medical Outcomes Partnership (OMOP), which draws on the resources of the pharmaceutical industry, academic institutions, nonprofit organizations, the FDA, and other federal agencies, may serve as a model of a viable public-private collaboration.\textsuperscript{33}

**Conclusions**

Secondary data, such as registries and retrospective databases, are often considered to complement randomized clinical trials since they are less costly and can be used to incorporate real-world experience to answer important health care questions. Analyses of secondary data provide a relatively efficient means for addressing hypotheses regarding adherence, treatment patterns, and burden of disease. However, effective use of secondary data requires addressing major methodological and infrastructural issues, including development of analytical tools to readily access and analyze data, formulation of guidelines to enhance quality and transparency, establishment of data standards, and creation of data warehouses that respect the privacy and confidentiality of patients. This paper described infrastructural requirements for secondary data utilization in the context of comparative effectiveness research and identified gaps that must be filled to address the underlying issues, with emphasis on data standards, data quality assurance, data warehouses, computing environment, and protection of privacy and confidentiality.
**Infrastructure Requirements for Secondary Data Sources in Comparative Effectiveness Research**

**Authors**

DEMISSIE ALEMAYEHU, PhD, is Executive Director, OR and Disease Area Statistics Statistic, and JACK MARDEKIAN, PhD, is Senior Director, OR Statistical Scientist, Pfizer Inc, New York, New York.

**DISCLOSURES**

This supplement was funded by Pfizer, Inc. Alemayehu and Mardekian are Pfizer employees. Alemayehu and Mardekian contributed equally to writing and revision of the manuscript.

**ACKNOWLEDGEMENTS**

Margaret McDonald, PhD, C. Daniel Mullins, PhD, Robert J. Sanchez, PhD, and Richard J. Wilke, PhD, provided valuable input to various draft versions of the manuscript.

**REFERENCES**


