Will the Growing Reliance on Real-World Data Fuel Fundamental Changes in the Way We Approach Database Analyses?

Stephanie Reisinger Vice President, Technology Solutions
Gary Schneider, MSPH, ScD Epidemiologist
Matthew Reynolds, PhD Vice President, Scientific Development

Electronic medical records and administrative claims databases, which contain "real-world" patient data collected at the point of care, have been used in pharmacoepidemiologic research for many decades. One of the first published database studies appeared in 1979, evaluating the association between the use of hormone replacement therapy in menopausal women and endometrial cancer using a database from the Group Health Cooperative (GHC) of Puget Sound.[1] Since that time, the focus of non-interventional research using real-world patient data has been relatively narrow, used mainly to fill information gaps not addressed through controlled clinical studies. However, the industry is currently in the middle of a fundamental shift in both the availability of, and reliance upon, real-world databases for evidence generation.

Several trends have converged to catalyze this shift including:

1. The demand for product value demonstration by an increasingly diverse group of stakeholders, including regulators and payers
2. Rapid proliferation, both in number and size, of available real-world data sources
3. Technological advances supporting the storage and management of "big data" assets
4. The development of specialized analytic methodologies to control for the types of bias found in real-world data sources[2]
5. A growing ability to support hybrid study designs, where patients analyzed retrospectively can be re-identified for prospective research

No longer just a sideline, the evidence generated from real-world data is rapidly becoming an integral component of new product evidence strategies.[3] At the same time, the growing volumes and heterogeneity of real-world data sources are creating analytic environments that are disorganized, inefficient and increasingly difficult to manage. Traditional database analytic approaches may be inadequate to fully take advantage of the evidence generation potential offered by this new era of real-world data.

Issues with traditional database analysis approaches in today's environment

Although most real-world databases contain similar information about patients collected at the point of care, these databases can vary significantly in both the structure and syntax of the data as well as the nomenclature used to represent pharmaceutical products and patient healthcare conditions. Because of these differences, the traditional analysis approach requires the development of a custom program written to answer a specific question against a specific database. This relies heavily on the availability of programmers with a sufficient understanding of the underlying database, a rate-limiting and inefficient approach that usually requires a single database to be selected for each study. This "one database per study" approach to evidence generation does not lend itself well to the growing demand for real-world evidence. Issues with the current approach include:

- **Not efficient:** Evidence generation is constrained by available programming resources and the knowledge of the programmers, and requires custom programming for each analysis.
- **Not transparent:** Patient and clinical event selection assumptions and algorithms are tied to the specific format of the database and embedded within the program code.
- **Not reproducible:** Format and programming differences among databases make it inefficient to execute and difficult to meaningfully compare evidence generated across disparate data sources.

Fueled by an increasing reliance on real-world evidence, pharmaceutical decision makers are demanding broader, more efficient evidence generation capabilities across heterogeneous real-world data sources, and new approaches are urgently needed to address this growing demand.

Standardization can help to address key issues

The issues described above are well known by most database researchers, and over the past seven years several organizations have focused on understanding and addressing them. In the United States, the Food and Drug Administration's (FDA) Sentinel Initiative,[4] the Observational Medical Outcomes Partnership (OMOP),[5] and the Observational Health Data Sciences and Informatics (OHDSI) collaborative,[6] and in Europe the EU-ADR project,[7] among others, all focus on the efficient use of real-world databases for evidence generation. A common theme across all these organizations is standardization, which falls into two broad categories: standardization of data and standardization of analytics.

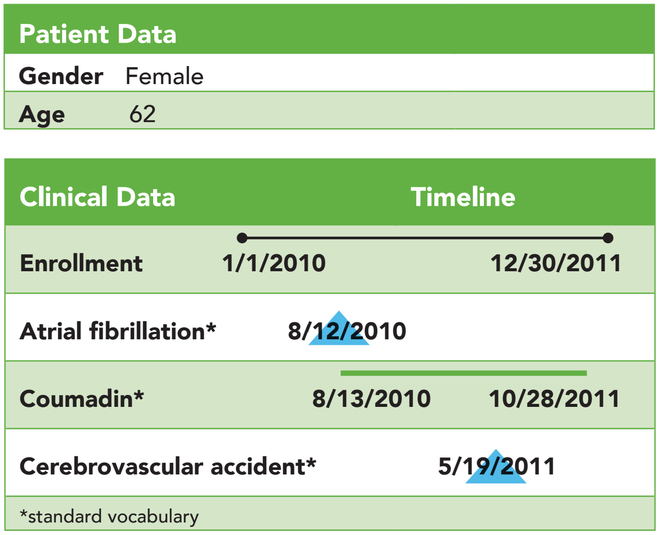
> **Fueled by an increasing reliance on real-world evidence, pharmaceutical decision makers are demanding broader, more efficient evidence generation capabilities across heterogeneous real-world data sources, and new approaches are urgently needed to address this growing demand.**

- ***Data standardization using a common data model:*** There have been several articles written about the development and use of a common data model (CDM) for analysis of real-world databases.[8,9] Although a CDM can be complex to implement, its basic purpose is fairly straightforward — to create a standard data format (structure and syntax) accommodating the critical data elements required to support the desired evidence generation capabilities efficiently. Some CDM designs, such as the OMOP CDM, also include a standardized vocabulary for drugs and conditions.[10]
- ***Analysis standardization using modular programs:*** A primary benefit of implementing a CDM is that standardized analytic routines can be written for the CDM and executed against any real-world database that has been transformed into the CDM format. Furthermore, key patient selection and analysis variables within each standardized module can be parameterized and entered by the user at analysis time. These "modular programs" can be executed by non-programmer researchers since they do not require any custom programming. Both the FDA's Sentinel Initiative and the OHDSI collaborative have included the development of parameter-driven modular programs as part of their respective research.[11,12]

A standardized analysis example

Below is a simplified illustration of a standardized analysis. Figure 1 provides a partial logical representation of a patient record in the OMOP CDM format, including demographic and clinical data. All the patient and clinical variables used in the example below are commonly available in real-world databases.

Figure 1: Partial patient record in OMOP CDM format



Patient Data	
Gender	Female
Age	62

Clinical Data	Timeline	
Enrollment	**1/1/2010**	**12/30/2011**
Atrial fibrillation*	**8/12/2010**	
Coumadin*	**8/13/2010**	**10/28/2011**
Cerebrovascular accident*	**5/19/2011**	

*standard vocabulary

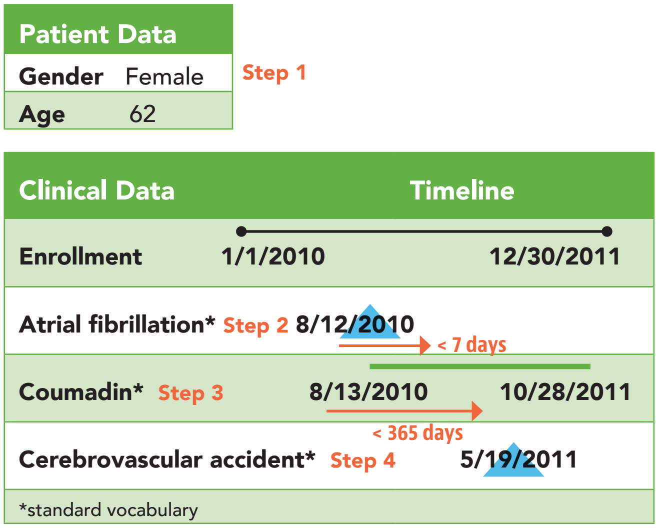
Table 1 and Figure 2 illustrate the steps and associated parameters required to perform a standardized analysis to answer a common type of analysis question.

Table 1: Example of modular program steps and associated user parameters

Analysis question:
How many female patients over age 60 who have been diagnosed with atrial fibrillation were also treated with Coumadin within 7 days of their diagnosis? Of those patients, what percentage had a stroke in the 365 days following diagnosis?

Modular Program Steps		User Parameters
Step 1	Select all patients with **user specified characteristics**	**Female; > Age 60**
Step 2	Restrict the patients selected above to only those patients with **user specified condition**	**Atrial fibrillation**
Step 3	Further restrict the selection to those patients who were treated with **user specified drug** within **user specified time frame**	**Coumadin; within 7 days after atrial fibrillation diagnosis**
Step 4	Of those patients, what percentage were diagnosed with **user specified condition** within **user specified time frame**	**Cerebrovascular accident; within 365 days after atrial fibrillation diagnosis**

Figure 2: Standardized analysis applied to a patient record in the CDM format



Patient Data		
Gender	Female	Step 1
Age	62	

Clinical Data		Timeline	
Enrollment		**1/1/2010**	**12/30/2011**
Atrial fibrillation*	Step 2	**8/12/2010**	< 7 days
Coumadin*	Step 3	**8/13/2010**	**10/28/2011**
		< 365 days	
Cerebrovascular accident*	Step 4	**5/19/2011**	

*standard vocabulary

Figure 2 provides a brief illustration of a standardized analysis, showing the analysis steps and how they are applied to the CDM. Although not appropriate for all analyses, there are many types of analyses that lend themselves well to this type of parameter-driven approach, including exploratory and descriptive analyses, analyses that are performed repeatedly (e.g., ongoing monitoring), common analytic calculations such as rates of diseases or outcomes, and characteristics of product exposure, to name a few.

"Collaborative analytics": A new era of real-world evidence generation

The potential for standardization to significantly improve the efficiency of real-world database analytics has been demonstrated through recent studies[13] as well as by research presented in this issue of *The Evidence Forum* in the article: *Collaborative Analytics in Action: A Case Study Focused on Treatment Patterns.* Yet there is another more subtle and potentially very powerful benefit of standardization that could fundamentally change the current database analytics paradigm.

Coding algorithms, which are defined as some combination of diagnosis, procedure, drug; or lab value codes and/or condition that reliably identify a specified health event from real-world databases, have recently received attention. Both the FDA Sentinel Initiative and the OMOP have published coding algorithms for various health outcomes of interest (HOIs) that are of particular interest to drug safety researchers.[14,15] Figure 3 shows an example coding algorithm for aplastic anemia. In an ideal world, all key clinical variables in a database study would be defined via coding algorithms, but in practice most of the algorithms required to identify clinical variables are custom developed (and redeveloped) for each study and database.[16]

Figure 3: Example coding algorithm for aplastic anemia from the OMOP HOI library

Example coding algorithm for Aplastic Anemia

ICD-9:284.0*, 284.8*, 284.9
AND within 60 days prior to the diagnostic code
Diagnostic procedure code for bone marrow aspiration or biopsy

In a standardized analytic environment such as the one described above, user parameters can be developed to standardize the implementation of coding algorithms for important clinical events. These parameters can be curated and stored in a clinical event library and later searched, shared, and re-used in analyses across an entire organization. Simply selecting the clinical event of interest from the library copies the appropriate parameters for that clinical event/coding algorithm into the desired analysis module.

The current analytic environment is based mainly on custom, one-off analysis programs developed in isolation against a single database for each study. Standardization enables an innovative environment of "collaborative analytics" where modular programs and clinical event definitions can be collaboratively developed, shared, and re-used within and across organizations and, thinking even bigger, across the entire industry. In addition, because modular programs and clinical event definitions can be executed against any data-source in CDM format, analyses can be efficiently reproduced across disparate databases and organizations, and the results of these analyses can be meaningfully compared.

Considerations and limitations

Although standardized analytics offers great potential to improve the power and efficiency of real-world evidence generation, there are some limitations to this approach.

- **Time and resource commitments:** The implementation of a CDM and a standardized analytic environment is complex and requires a commitment of time and resources.
- **Information loss:** The process of mapping the raw source data into the CDM may result in some data loss, particularly if non-standard drug and condition codes are found within the source data. To mitigate this issue, some CDMs, such as OMOP, allow the native codes to be stored and used for analysis in addition to the standardized vocabulary.
- **Clinical and data content expertise:** Standardization does not reduce the need to have clinical, epidemiological, and data content experts involved in the development of study protocols and analysis parameters and for interpretation of results.
- **Interoperability:** Not all types of analysis are well suited for standardization. Organizations will continue to have the need for custom analysis programs to be written for detailed and difficult analytic tasks. Interoperability between the standardized and traditional analytic environments is necessary for researchers to move back and forth between environments.
- **Quality of Output:** Standardized analytics are powerful and efficient, creating an environment with a potential for misuse by untrained and inexpert users. Formal user training requirements, access limitations, and peer review processes should be developed and implemented to ensure analysis results are of the highest quality.

To reach the full potential that standardization can provide, the industry should consider moving toward the adoption of an industrywide common data model standard for real-world analytics.

Where do we go from here?

Standardized analytics offers great potential to address growing demands for efficient real-world evidence generation, but we are only at the beginning of our understanding of how to best integrate this approach into existing evidence generation schemes. To reach the full potential that standardization can provide, the industry should consider moving toward the adoption of an industrywide common data model standard for real-world analytics. Given that there are multiple organizations promoting different CDM versions, this statement may seem controversial. However, existing CDM standards proposed by different organizations are more similar then they are different, and recent research has provided insight into the pros and cons of each model.[17] An ideal standard would incorporate the best features of each.

Moving forward, collaborative research organizations such as OHDSI are critical in providing a platform to advance the science of standardized analytics while integrating the input of diverse stakeholders. Finally, commercial technology and data providers should incorporate non-proprietary, open standards into their offerings where commercially feasible, ensuring greater interoperability and integration across all commercial real-world data offerings.

For more information, please contact Stephanie.Reisinger@evidera.com, Gary.Schneider@evidera.com or Matthew.Reynolds@evidera.com.

REFERENCES ON NEXT PAGE

REFERENCES

[1] Jick H, Watkins RN, Hunter JR, Dinan BJ, Madsen S, Rothman KJ, et al. Replacement Estrogens and Endometrial Cancer. *N Engl J Med.* 1979 Feb; 300(5):218–222.

[2] 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 Jan; 21 Suppl 1:41–49.

[3] Thwaites R. So What Exactly is Your Real-World Data Strategy? *The Evidence Forum,* March 2014. Available at: http://www.evidera.com/wp-content/uploads/2014/03/The-Evidence-Forum-2014-March.pdf. Accessed March 29, 2015.

[4] FDA's Sentinel Initiative. Mini-Sentinel Website. Available at: http://minisentinel.org. Accessed March 29, 2015.

[5] Observational Medical Outcomes Partnership (OMOP). Observational Medical Outcomes Partnership Website. Available at: http://omop.org. Accessed March 29, 2015.

[6] Observational Health Data Sciences and Informatics (OHDSI) Website. Available at: http://www.ohdsi.org. Accessed March 29, 2015.

[7] EU-ADR. EU-Adverse Drug Reactions Website. Available at: http://euadr-project.org. Accessed March 29, 2015.

[8] Reisinger SJ, Ryan PB, O'Hara DJ, et al. Development and Evaluation of a Common Data Model Enabling Active Drug Safety Surveillance Using Disparate Healthcare Databases. *JAMIA* 2010 Nov-Dec; 17(6):652–662.

[9] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a Common Data Model for Active Safety Surveillance Research. *JAMIA.* 2012 Jan-Feb; 19(1):54–60.

[10] Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of Alternative Standardized Terminologies for Medical Conditions within a Network of Observational Healthcare Databases. *J Biomed Inform.* 2012 Aug; 45(4):689–696.

[11] FDA Sentinel Routine Querying Tools. Mini-Sentinel Website. Available at: http://www.mini-sentinel.org/data_activities/modular_programs/default.aspx. Accessed March 29, 2015.

[12] OHDSI Analytic Tools. OHDSI Website. Available at: http://www.ohdsi.org/analytic-tools/. Accessed March 29, 2015.

[13] Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, DeFalco FJ, et al. Feasibility and Utility of Applications of the Common Data Model to Multiple, Disparate Observational Health Databases. *JAMIA* 2015 Feb 10; 02:1–12. doi:10.1093/jamia/ocu023.

[14] OMOP HOI Library. OMOP Website. Available at: http://omop.org/HOI. Accessed March 31, 2015.

[15] FDA Sentinel Identification of Health Outcomes. Mini-Sentinel Website. Available at: http://minisentinel.org/methods/outcome_identification/default.aspx. Accessed March 31, 2015.

[16] Reynolds M. Increasing the Value of Database Research with Validated Coding Algorithms. *The Evidence Forum,* March 2014. Available at: http://www.evidera.com/wp-content/uploads/2014/03/The-Evidence-Forum-2014-March.pdf. Accessed March 29, 2015.

[17] Kahn MG, Batson D, Schilling LM. Data Model Considerations for Clinical Effectiveness Researchers. *Med Care.* 2012 Jul; 50 Suppl:S60-67.

Projected sales of main products in 2013
marketing participation in the securities market
Revenue growth divisions
TYU division
ERT division
Share of market activity
Distribution of the securities market key players
Distribution marketing participation in the securities market