

Increasing the Value of Database Research with Validated Coding Algorithms

By Matthew W. Reynolds, PhD, Vice President, Epidemiology

The employment of high quality methods for retrospective database research has never been more clearly relevant and important given today's pharmaceutical research environment. An analysis from 2010 shows a significant increase over time for claims database (see Figure 1) and electronic medical record (EMR) database studies (see Figure 2). Although the analysis has not been updated with data from 2011 to the present, the expanding focus on the use of retrospective databases suggests the continuation of this exponential increase. We have seen a tremendous effort in the design of large safety-based database initiatives in the United States (e.g., Sentinel Initiative) and Europe (e.g., EU-ADR), as well as the creation of groups like the Observational Medical Outcomes Partnership (OMOP) which have focused many of their efforts in the area of epidemiological and safety-based database research methods. One of the areas of particular interest for these groups has been the identification and development of validated coding algorithms for use in identifying and defining study cohorts, health outcomes of interest, as well as patient comorbidities. Validated coding algorithms are important for a couple of reasons. First, the process of defining them has not always been as rigorous as desired (or well published), leading

to incorrect identification of patients, misclassification of events and costs, and inaccurate research results. Secondly, there is an opportunity to leverage the expansive number of databases available today, with access to tens and sometimes hundreds of millions of patients, to meet regulatory requirements. Increasingly, regulatory agencies in both the U.S. and Europe are allowing the use of these databases in a rolling retrospective way to meet post-marketing commitments that historically would have had to be done through registries or chart reviews.

A coding algorithm can be defined as a combination of diagnosis, procedure, drug, or lab value codes (e.g., ICD-9, CPT-4, NDC) and/or conditions (e.g., diagnostic code in the primary position of a hospital claim, minimum length of stay in a specific care setting) that can be used to identify a specific clinical term in an electronic healthcare database. Hence, all key clinical variables in a database study would be defined via coding algorithms. Some may be simple (e.g., a single diagnosis code), while others could prove notably more complex (e.g., a diagnosis code in a primary hospital position within 30 days of a second diagnosis code). All these clinical variables would be expected to be defined and operationalized prior to conducting the associated database analyses.

THE OPTIMAL APPROACH FOR CONDUCTING HIGH QUALITY DATABASE RESEARCH WOULD INCLUDE THE IDENTIFICATION, ASSESSMENT, AND INCORPORATION OF VALIDATED CODING ALGORITHMS INTO THESE DATABASE STUDIES.

In most publications using claims or EMR databases, the authors rarely provide full descriptive definitions for how the key variables were operationalized and how those definitions were determined. It is more common that an author may provide some definition for the study cohort of interest and/or the key health outcome of interest, but rarely are other clinical covariates ever defined. Further, while it may be more common to define the cohorts and health outcomes of interest, it is rare that the authors note how and

CLAIMS DATABASE PUBLICATIONS

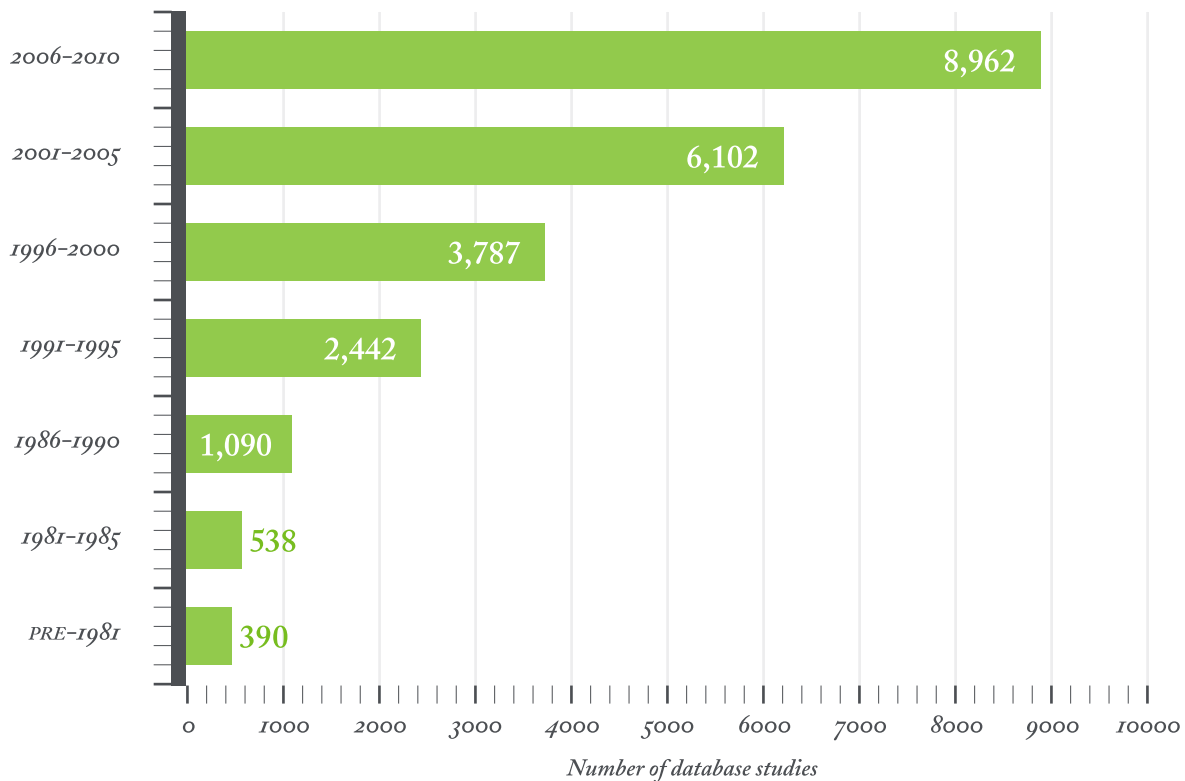


figure 1

why the definitions were determined or developed. This becomes a major issue when attempting to determine how to compare results from multiple database studies that clearly utilized different definitions for their key variables and/or never defined their key variables at all. While many databases clearly require variations in coding due to the inherent differences in the underlying coding systems (e.g., ICD-9 vs. OXMIS), the coding variations across studies limit our ability to quickly detect important patterns in the natural history of disease, and they further impede our ability to reach defensible conclusions about the safety, effectiveness, and cost-effectiveness of existing treatments.

The optimal approach for conducting high quality database research would

include the identification, assessment, and incorporation of validated coding algorithms into these database studies. Employing knowledge from similar, published database studies that demonstrate the effectiveness, or lack of effectiveness, of various coding algorithms for specific clinical events would help to assure that database studies are accurately and completely identifying the appropriate clinical events of interest. Using information about the positive predictive value (PPV – the proportion of positive results that are true positives), sensitivity (the percentage of people correctly identified as having the condition being studied), and specificity (the percentage of people correctly identified as not having the condition being studied) of various coding algorithms would

help to drive the decisions about how best to define the clinical events of interest in each database study. Ideally, a score of 75% or higher in all three areas is desired for optimal results. To highlight this importance, in August 2010, the Database Special Interest Group for the International Society of Pharmacoepidemiology (ISPE) conducted a workshop to provide guidance to database researchers regarding the identification, development, validation and translation of coding algorithms in electronic healthcare databases.

The literature is the first place to start when identifying the best coding algorithm to use when defining clinical terms of interest. A very clear list of clinical terms should be created, including disease, terms of interest

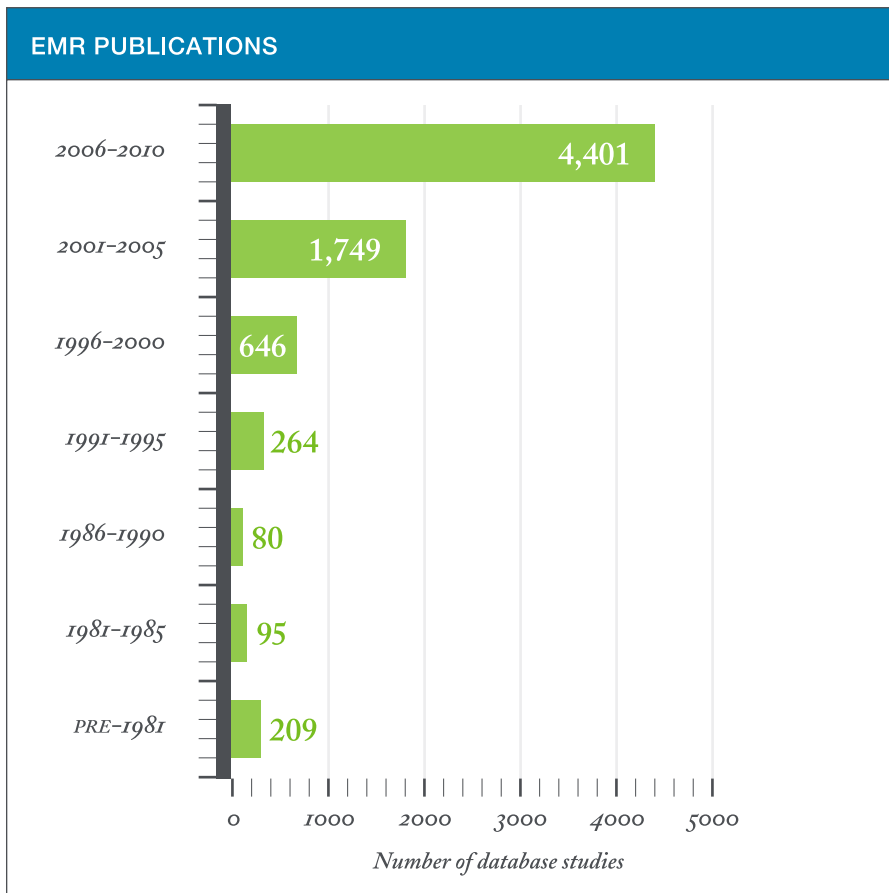



figure 2

and study types, and it is important to be as specific as possible since the coding algorithms that are built are typically going to tie to a very specific clinical event of interest. Focusing on claims and EMR databases will help narrow the search strategy, however, this can be challenging since EMBASE and MedLine only recently established good search terms for databases and PubMed still has not. Limits and criteria need to be thought through very carefully. When considering how far back to look, consider things such as how valid an algorithm from 10 years ago would be now, or have treatment patterns and definitions changed? It is also important to note that peer-reviewed publications may not be plentiful in this area, so conference abstracts should also

be considered since better results may be found here as opposed to published articles alone. A screening strategy then needs to be developed to identify which studies should be used and which should not. Typically, any database study that has the specified clinical term of interest with clear definitions would be kept and then prioritized. The best studies are those that have used the database of interest and include a detailed coding strategy along with validation metrics. References from publications can also be explored to further expand the possibility of viable studies. Contacting authors directly is another option to identify codes used in previous studies.

When there is nothing in the literature, validated coding algorithms need

to be developed from scratch. Past studies from the literature can be assessed to see which codes were used, even if they were not validated. Medical coders can provide insight into which codes are typically submitted for reimbursement for specific diseases and treatments. Clinicians can provide valuable insights, such as how commonly they use particular codes in their practice. The clinical insight is invaluable to better understand the patient evaluation, diagnosis, referral, and treatment patterns which will drive the engineering of optimal coding algorithms. Knowing factors such as the place of service, the physician type and timing between codes/visits can be instrumental in the building of coding algorithms.




Given the difficulty in identifying and synthesizing this evidence, combined with a desire to ensure consistency of definitions across studies, some industry groups, such as the Pharmacoepidemiology and Database Research Unit at Merck and Company, have developed a central coding library based on literature and clinical expertise to support their clinical and epidemiology research needs. Groups like OMOP have explored a variety of ways to best identify all available information on coding algorithms and how best to employ them consistently across databases.¹

Validation of an algorithm can be very complex, but basically, once an algorithm is built, it needs to be shown that it really works. Does it actually

identify the patients needed? Does it discriminate between cases and non-cases? Validation requires a gold standard to define the case. Most published studies have required chart reviews, but more commonly, EMR databases can also be used to create an algorithm based on components from both billing interactions and clinical chart/text information. The key is being able to reliably identify true cases and non-cases to build a validated coding algorithm. Think through and analyze your algorithm and then apply it. Calculate the PPV, sensitivity, and specificity to see if any modifications to the algorithm are needed based on those results.

Validated coding algorithms provide quality, reliable definitions for diseases,

comorbidities and clinical endpoints, and when well defined and able to be referenced, they strengthen the quality, value, credibility, and replicability of studies. They produce better study results compared to those that may be using imprecise definitions, an absolute necessity in the future for studies being used for regulatory and reimbursement agencies. Organizations can provide consistency of definitions across studies by building a library of validated coding algorithms and appropriate definitions that reflect the clinical events being studied. By referencing them in peer-reviewed publications and providing transparency in database studies, the entire research community is served. 

For more information, contact Matthew.Reynolds@evidera.com.

References

¹ LoCasale B, Mehta V, Alderton L, Bortnichak E, Reynolds M, Reejis S, Jones N. Systematic Process for Coding Definition/Algorithm Development in Database Analyses. 25th International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Providence, Rhode Island, August 16–19, 2009.