Outcomes Research and Machine Learning: Complementary or Incompatible?

Lesley H. Curtis, PhD, Professor in Medicine, Duke University School of Medicine and Duke Clinical Research Institute, Durham, NC, USA

KEY POINTS . . .

Defining and refining the research question is an essential element of any scientific inquiry and applies equally to outcomes research and machine learning.

Preparation of data for analysis—whether via traditional outcomes research methods or with machine learning approaches—requires a solid understanding of the work flow that generated the data.

The critical elements of rigorous outcomes research are directly transferrable to a variety of methodological approaches including machine learning.

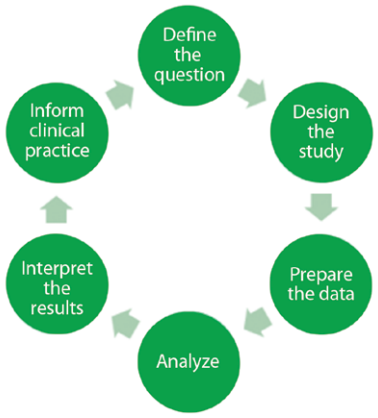
For additional information in this issue:
"Big Data" was the topic presented at the **ISPOR 20th Annual International Meeting** in Philadelphia, PA, USA, as part of the Third Plenary Session: "Big Data, Big Systems, and Better Evidence: What Progress?" (see page 24 ⇨).

The following article was based on a presentation given during the First Plenary Session at the ISPOR 19th Annual International Meeting, May 31- June 4, 2014, Montreal, QC, USA

The outcomes research process begins by defining the question to be answered. (Fig. 1) According to noted statistician John Tukey [1], "finding the question is often far more important than finding the answer." This same notion of exploratory analysis can be applied to how one thinks about machine learning and other types of approaches, where computational power bears on data in order to generate new knowledge. Typically, the exploratory analyses done by outcomes researchers are guided at the very least by a kernel of an idea that is either known or believed, in conjunction with some existing data. Additionally, the question must be grounded in clinical practice or in policy; in particular, a focus on the centrality of the patient and provider will help to define the relevant questions that need to be answered. In other words, the data informs the question being asked.

Once the question has become well-defined (including being considered relevant to the providers and decision makers), and the appropriate data has been assembled, the second step begins: designing the study. The study design must encompass concepts such as: who is to be included and excluded? How will the outcomes of interest and co-variates of interest be measured? What analytic methods should be used? As said by Cook and Steiner [2], "it's not possible to put right with statistics what you've done wrong by study design." Within the context of machine learning, this idea can be adapted further to warn that it is also not possible to put right with computational power, what has been done wrong by study design. Thus, one of the first questions that must be asked in designing a study is: are the data that we have fit for the purpose to which we tend to applying them? The suitability or "fit-ability" of the data must therefore be considered.

Figure 1: The Outcomes Research Process.

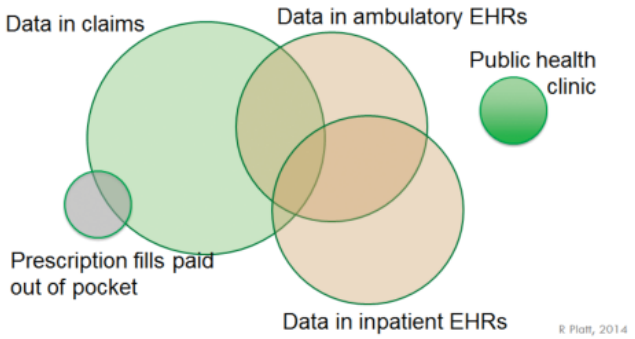


There are a few concepts to keep in mind during this stage. Health care is fragmented, and one person's data may exist in many different places. Depending on a variety of factors, there is likely some overlap between the rich data in ambulatory electronic health record systems and inpatient data (i.e., data from inpatient electronic health record systems), though as shown in Figure 2, seldom is there a complete overlap. Claims bring an important dimension to this in terms of ability to ascertain person-time—the time(s) during which observation is complete. In ambulatory electronic health records and inpatient electronic health records, knowledge can be gleaned on who has been seen, but the absence of the care is not particularly meaningful. This is due to the fact that care could have been delivered in a setting that is not part of the system included in the electronic health records, and so there would be a gap in information for that encounter. Depending on the population of interest, there may be no ability to track the data, such as in the case of out of pocket prescription payments, and individuals who receive care in settings not reflected in the electronic health record. To further complicate things, the environment is also in a state of perpetual change and evolution –providers as well as locations, departments, and clinics all come and go; applications to capture data also may change the context in which data are collected.

What is meant by "complete data" is another important concept to consider. Completeness might depend on a variety of factors, including: the specific conditions, treatments, and outcomes of interest, and identifying to what extent inpatient and outpatient electronic health record data are already integrated and potentially integrated with claims data. Additionally, one must consider the structure of the health system, the level of insurance coverage – a key determinant, the geographic region within the health system, and referral patterns. As an example, if a person received a diagnosis or underwent a treatment during a time period, would we observe it? And in what data source? There needs to be an affirmative answer to the questions that arise, and the source of the data must also be identified (Fig. 1). It is imperative that as the study is being designed, the researcher is taking into consideration these aforementioned issues.

Yet another issue that outcome researchers should be increasingly thinking about is data standardization across organizations. A recent paper published by Marsha Raebel and her Mini-Sentinel colleagues describes the lack of standardization in hemoglobin A1c (HbA1C) and platelet units across twelve data partners [3], and 34 different result units for HbA1C and 68 different result units for platelets. It might be intuitive for a human to go through and clean up, or standardize, the information, but a machine would need to be taught to perform that function. This aspect is a critical piece of preparing any data set for analysis, and shows the necessity for that human component in machine learning methods.

Figure 2: One person's data may be in different places.



With the question and study design in place, the data set must be deemed sufficient enough to be able to answer the question at hand. The preparation of data for analysis is the third key element in the outcomes research process. A primary concern during this step is having a good understanding of the work flow that generated the data being used. Whether talking about electronic health record data or claims data, the perspective from which those data come is critical and will guide our interpretation. Additionally, the data partners and experts who understand the data must be engaged in this process and regarded as a resource to gain insight into the clinical/general work flow that created the data. Lastly, data quality problems must be expected, and in turn acknowledged, evaluated, and addressed. Often these problems come in terms of missingness or inconsistencies, and a data set without missing data and/or inconsistencies should be considered suspect.

... the outcomes researcher and machine learning (or any other kind of data mining type of approach) should be thought of as not only compatible, but complementary to one another.

Once the data has moved out of the preparation phase, it must be analyzed and then interpreted. It is very important that the interpretation is done with caution, as there may be a number of complicating issues. To begin with, the data may not represent a random sample of patients, and the desired granularity or the completeness and quality of data may not be present; within the context of machine learning-type approaches, the reference standards that are used for prediction might not be adequate. It is important to also keep in mind that the measurement of concepts, coding, etc. changes over time. The big changes, ICD-9 to ICD-10, are commonly thought of, but there are small and important changes that happen as well: providers leave data sets, clinics drop out of EHR systems, etc. Unless there is ample understanding of what has changed, the interpretation of the findings may not be accomplished in an accurate way.

The last step, and the ultimate goal of the outcomes researcher, is to inform clinical practice and eventually improve care. This involves asking the right question, choosing the optimal approach, knowing one's data, preparing that data for analysis, and then answering in a meaningful and actionable way. Outcomes researchers bring the quintessential skill set, knowledge, and necessary expertise to the table in using these approaches, so it can be concluded that the outcomes researcher and machine learning (or any other kind of data mining type of approach) should be thought of as not only compatible, but complementary to one another.

REFERENCES

[1] Tukey JW. We need both exploratory and confirmatory Am Stat 1980;3423-5. [2] Steiner PM, Cook TD, Shadish WR, Clark MH. The importance of covariate selection in controlling for selection bias in observational studies. Psychol Methods 2010;15:250-67. [3] Raebel MA, Haynes K, Woodworth TS, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. Pharmacoepidemiol Drug Saf 2014;23:609-18. ■

Additional information:
To view Dr. Curtis' presentation, please visit the Released Presentations page for the 19th Annual International Meeting at http://www.ispor.org/Event/EventInformation/2014Montreal?p=212