

Clinical Data Repositories:

Less Than Meets the Eye

There has long been a belief that clinical data repositories are potential gold mines of untapped knowledge, and that with the appropriate electronic infrastructure, they would serve as a source of new information about the causes of disease, the identity of new biomarkers, and other unappreciated statistical correlations. Using the awesome power of modern data

because patients insist on the tests, and many other spurious reasons. These sources of bias, importantly, are typically not recorded anywhere in the medical record, and so the task of statistically untangling the data in order to



We should try our best to learn new things with the data, but our expectations should be low. More importantly, our skepticism about whatever is discovered should be very high.

mining and machine learning methods, we would be able to troll years of clinical data and extract gold. I am afraid that this view is overly optimistic.

First, clinical data repositories are the historical record of physicians and other healthcare providers ordering tests, procedures, and documenting their inference in an extremely biased manner. The goal is not to objectively sample reality, but to build a story that convinces themselves and others that they have the right model of what is wrong with the patient, and that their actions are reasonable. This is not a bad thing—this is the exercise of the art of medicine (which is still very much an art, despite our attempts to codify and standardize)—but it is not a good basis upon which to build a discovery engine. Second, clinical records were not invented to support research and many of the elements required for good research are not present. All tests must be interpreted with knowledge of the prevalence as well as the sensitivity and specificity. Thus, taken out of context, the results of a test are very difficult to interpret with respect to their

accuracy and information content. Physicians also sometimes order tests for medico-legal reasons, financial reasons, to document the course of a disease,

generate a clean and believable dataset is incredibly difficult. Third, the practice of medicine and the use of medical tests, procedures and terminology are constantly evolving, and thus any attempt to combine data over a significant period of time (even as short as a few years) is likely to be confounded by changes in practice, the arrival of new therapeutic and diagnostic capabilities, and simple “medical fashion.”

Having expressed my pessimism, I believe we still should apply data mining algorithms to these data, and attempt to overcome these difficult challenges. In our initial investigations, we are likely to discover the obvious: typical patterns of test ordering associated with medical practice, and the trivial correlations between different data sources in the record. We should try our best to learn new things with the data, but our expectations should be low. More importantly, our skepticism about whatever is discovered should be very high. Data mining activities may suggest new hypotheses, and these can then be followed up with careful analysis of clean (ideally prospective) data. But to expect a goldmine of discoveries, at least today, is to underestimate the difficulty in preparing this data for serious use in discovery.

DETAILS

Russ B. Altman, MD, PhD, is principal investigator for Simbios, a National Center for Biomedical Computing, and professor of bioengineering, genetics and medicine at Stanford University.

SEND US IDEAS Got your own opinions on this topic? Or have another topic you'd like to write about for these pages? Send us your thoughts on the Feedback page of our Web site: <http://biomedicalcomputationreview.org/feedback.html>.



More Valuable Than You'd Expect

The recent federal stimulus package is injecting billions of dollars into electronic health record implementation. To determine whether such records will be useful for clinical research in the genomic era, three questions should be answered: 1) As compared to what? 2) To do what? 3) At what cost in time and treasure?

With regard to the first question, we

currently go unrecognized in our health care system—often until it is too late. A prime case of this is the very large number of deaths attributable to Vioxx. Similarly important but rarer events such as the pancreatitis associated with xenatide can be sussed out even if the exact magnitude of the effect is in question. Such findings would then trigger further, perhaps better-controlled, studies. In addition, when a study would require hundreds of thousands of patients to measure a low-magnitude effect, electronic health records provide a unique resource, particularly when they are carefully mined using natural language processing techniques. Indeed much of the purely claims-based research is vulnerable to both the

Given the availability of clinical data obtained from our very expensive and intensive health care process, we must at least determine the extent to which electronic health record information can further science and improve diagnostic and treatment modalities.

know that even highly regarded large cohort studies such as those published regularly in the genome-wide association studies literature are highly prone to phenotypic misclassification. We also know that carefully selected populations exhibit different characteristics than the population of health-care recipients as a whole. For example, populations selected for a study may not manifest all the risks and interactions of exposure to a particular therapeutic drug. Moreover, when human beings extract medical records or ask questions, they inject variability and biases, evident in a review of any of the annotations in existing studies. In contrast, we can repeatedly run the entire corpus of suitably de-identified clinical records through different natural language processing methods and filters of varying stringency, and compare patient characteristics at one hospital to those at another. We can do so comprehensively, repeatedly, and have available large numbers of controls for confounding factors. Moreover, whereas classical recruited cohort studies usually face significant challenges in obtaining adequate representation of underrepresented minorities, comprehensive electronic medical records typically include more members of these groups.

With regard to the second question, there is no doubt one has to be sober and careful regarding what kinds of questions can be answered using data from electronic medical systems. Nevertheless, there are numerous “low hanging fruit” that enable electronic-health-record-based research projects to proceed productively. For example, any methodological and timely review of health-care system data could identify certain noteworthy high-magnitude epidemiological effects that

coarseness and reimbursement bias of the characterization provided by billing codes (e.g., a radiologist coding a “rule out rheumatoid arthritis” x-ray with the diagnosis of rheumatoid arthritis even if the patient does not have such disease).

And as for the cost in time and treasure, if we can conduct *in silico* observational studies in one hundredth of the time and for one tenth, hundredth or even thousandth of the cost of a conventional observational study, we should do so—with appropriate adjustments for bias, variation, and multiple hypothesis testing. We owe it to the public to at least explore what the results might be, especially if they identify dangerous drugs or promising new therapies. Given the availability of clinical data obtained from our very expensive and intensive health care process, we must at least determine the extent to which electronic health record information can further science and improve modes of diagnosis and treatment.

There is, of course, no doubt that many studies are answerable only through classically organized randomized controlled trials or tightly selected observational studies. However, the promise of electronic health records was never that they would be the only platform on which clinical research could be conducted in the future but “merely” an important component of the research agenda at the national and international level. □

DETAILS

Isaac Kohane, MD, PhD, is the principal investigator for Informatics for Integrating Biology and the Bedside (i2b2), as well as Lawrence J. Henderson Associate Professor of Pediatrics and Health Sciences and Technology at Harvard Medical School, and Chair of the Informatics Program at Children's Hospital, Boston.