Using “Big Data” to Dissect Clinical Heterogeneity

Russ B. Altman, MD, PhD; Euan A. Ashley, MRCP, DPhil

Experienced practitioners recognize that patients with a shared diagnosis often fall into subsets that “look” the same and often respond to similar treatment strategies. Indeed, the source of clinical wisdom within many specialties is a nuanced knowledge of these subsets; they typically are not described in standard texts and may be recognized only after years of delivering care. The best clinical teachers convey these distinctions to their trainees, thus giving the trainees knowledge that might otherwise be acquired only after observing hundreds or thousands of clinical cases.

One of the promises of “big data in medicine” is to accelerate our ability to recognize disease heterogeneity and to create new distinctions using large numbers of measurements on large populations of patients. The opportunities for big data have exploded with improvements in our ability to measure biomarkers (clinical laboratory, genetic, proteomic, metabolomic measurements), to collect images, to instrument patients with personal devices, and to store all these data in electronic formats. The availability of these data suggests that we may be able to dissect clinical heterogeneity on the basis of not just clinical observations on our own patients but also aggregated measurements in and observations about large populations of patients. It is in this context that the work of Shah and coworkers in this issue of Circulation sought to understand clinical subtypes of heart failure with preserved ejection fraction (HFpEF). Admittedly, their data set is not comparable to those heterogroups of cohorts and signals of efficacy are reduced or lost. Indeed, awareness of the heterogeneity of many common and complex diseases led the Institute of Medicine to call for a re-examination of the way we define diseases, taking advantage of our ability to measure disease at the molecular, cellular, tissue, and whole-organism levels.

One of the promises of “big data in medicine” is to accelerate our ability to recognize disease heterogeneity and to create new distinctions using large numbers of measurements on large populations of patients. The opportunities for big data have exploded with improvements in our ability to measure biomarkers (clinical laboratory, genetic, proteomic, metabolomic measurements), to collect images, to instrument patients with personal devices, and to store all these data in electronic formats. The availability of these data suggests that we may be able to dissect clinical heterogeneity on the basis of not just clinical observations on our own patients but also aggregated measurements in and observations about large populations of patients. It is in this context that the work of Shah and coworkers in this issue of Circulation sought to understand clinical subtypes of heart failure with preserved ejection fraction (HFpEF). Admittedly, their data set is not comparable to those associated with Internet commerce or even the human genome.

These disease subsets often turn out to be physiologically distinct processes that simply share a common clinical end point. Failure to recognize these subsets can confound treatment decisions and clinical research studies in which treatments are applied to heterogeneous cohorts and signals of efficacy are reduced or lost. Indeed, awareness of the heterogeneity of many common and complex diseases led the Institute of Medicine to call for a re-examination of the way we define diseases, taking advantage of our ability to measure disease at the molecular, cellular, tissue, and whole-organism levels.

One of the promises of “big data in medicine” is to accelerate our ability to recognize disease heterogeneity and to create new distinctions using large numbers of measurements on large populations of patients. The opportunities for big data have exploded with improvements in our ability to measure biomarkers (clinical laboratory, genetic, proteomic, metabolomic measurements), to collect images, to instrument patients with personal devices, and to store all these data in electronic formats. The availability of these data suggests that we may be able to dissect clinical heterogeneity on the basis of not just clinical observations on our own patients but also aggregated measurements in and observations about large populations of patients. It is in this context that the work of Shah and coworkers in this issue of Circulation sought to understand clinical subtypes of heart failure with preserved ejection fraction (HFpEF). Admittedly, their data set is not comparable to those associated with Internet commerce or even the human genome.

However, 397 patients with 46 measurements provide 18 262 individual data points that are impossible for an individual clinician to analyze fully. HFpEF represents a challenging clinical problem with few effective treatments. Unlike heart failure with reduced ejection fraction, for which outpatient treatments have evolved and improved over the last decade, HFpEF has been particularly resistant to improvement in outcomes. It thus represents an attractive target for the tools of modern big data and machine learning, tools that can find patterns in complex data and perhaps define subsets that are more cohesive.

Shah and coworkers gathered 397 cases of HFpEF and used 46 continuous phenotypic variables comprising clinical, laboratory, electrocardiographic, and echocardiographic findings to define mutually exclusive groups of patients whose within-group variability is much less than the variability between groups. This activity is called phenomapping because it creates a landscape of HFpEF patients based on their aggregated phenotypic features. For this study, the data collected about these patients was exquisitely detailed, including even noninvasive measures of pressure and volume. After applying their computational methods to the data, the authors defined 3 clinical syndromes with satisfying features: They differ along sensible clinical parameters, and they are predictive of outcome severity. Importantly, a new set of cases (not used in defining the 3 syndromes but selected from the same institution with the same criteria) allowed the authors to test the local generalizability of their results and to show that the new cases have outcomes that are reasonably well predicted by the syndromes into which they are classified.

It is particularly satisfying that complex machine learning algorithms wind up defining clinical syndromes that can be described, at least at a high level, with the same sort of language that an experienced physician might use. They include the young patient with normal B-type natriuretic peptide and diastolic dysfunction; the obese, diabetic, apneic patient with poor left ventricular relaxation; and the older hypertensive, renal-compromised patient with extensive remodeling. Of course, these summaries hide the fascinating and complex mathematical interactions between the phenotypic variables that define the syndromes. Nonetheless, they provide an intuition about these patients that makes clinical sense, and more important, they correlate with outcomes of cardiovascular hospitalization and death.

Readers of Circulation may not be accustomed to articles with the machine learning approaches used by Shah et al, and the clinical relevance of the results is the most important finding of the paper. Nevertheless, the article provides an excellent tutorial on some of the issues that must be examined closely in evaluations of this type of work. The key issues include the following:

1. Definition of the cases (and lack of controls). The authors explain their criteria for defining cases of HFpEF, and
this is critical. Their study lacks controls; patients without HFpEF are not used to define the clusters of HFpEF. For this reason, we cannot speak confidently about the relevance of the discovered clusters in the larger population. Indeed, the authors make no claims in this regard. If patients have a diagnosis of HFpEF (consistent with that used by the authors), then these 3 syndromes are relevant. However, these syndromes are not useful for finding HFpEF cases because we do not know the prevalence of similar phenotypes in the overall population.

2. Missing data. It is not uncommon for data sets to have some values missing because they are unmeasured or unrecorded. The authors used a strategy of imputation in which they predicted reasonable values for missing data to allow more cases to be included in the analysis. Of course, these estimated values must be evaluated at the end of the analysis to ensure that they did not drive the principal results.

3. Correlated features. The authors evaluated 67 total variables to characterize patients and settled on 46 after measuring the levels of correlation (and missingness) in the original 67. Machine learning methods have different performance profiles in the setting of data that is redundant; some can be misled by the redundancy and may essentially double-count evidence, whereas others are more robust to correlated data.

4. Choice of similarity metric. The authors use simple euclidean distance between patients (who are described as a vector of 47 numbers), a reasonable choice, but there are many other metrics of similarity that can yield different qualitative and quantitative results.

5. Use of area under the receiver-operating characteristic curve. This analysis allows the fundamental tradeoff between sensitivity and specificity to be finessed by examining performance of a classifier over all potential cutoffs, from liberal (higher sensitivity at the expense of more false positives) to strict (higher specificity but more false negatives). Their method achieves a good cutoffs, from liberal (higher sensitivity at the expense of more false positives) to strict (higher specificity but more false negatives). Their method achieves a good area under the receiver-operating characteristic curve but by no means has definitive predictive performance. To be fair, the authors stress that prediction (and competition with MAGGIC [Meta-Analysis Global Group in Chronic Heart Failure] and other metrics) was not the goal of the work.

6. Replication and separation of training and validation sets. It is axiomatic in machine learning that the data used to create a model cannot be used to validate it. The phenomenon of overfitting occurs when a classifier works very well on the cases it has seen in training at the expense of very poor performance on new, unseen cases. There are several ways to evaluate the potential for overfitting, but the best is probably a “held-out” validation set of new cases with which the classifier performance is evaluated. Ideally, the performance on new cases should be comparable to the performance on the training cases. In this work, the authors tested their method on a validation set comprising a prospective cohort of 107 individuals from the same institution.

Experts in machine learning can debate the specific approaches taken and the technical advantages of each. However, the choices of Shah and coworkers are generally reasonable and lead to results that provide clinical insights. The authors’ claim that the 3 groups are mutually exclusive is understandable but may be overstated; a larger sample of patients may begin to show populations of patients falling between the defined clusters, thus making their absolute boundaries somewhat unclear. At this point, however, we have not found such patients. The distinction between HFpEF and heart failure with reduced ejection fraction itself is something that the authors take as a given but is defined almost entirely by a single phenotype: the ejection fraction. It would be fascinating (albeit challenging) to use similar methods on the entire population of heart failure patients to see if the HFpEF and heart failure with reduced ejection fraction distinctions stand up to the scrutiny of big data.

In the end, we are left with a modern analysis of a heterogeneous disease and a credible trio of syndromes that have demonstrably different outcomes: the young patient; the obese, diabetic patient; and the older patient. These may be useful in the design of future trials for HFpEF—or the retrospective reanalysis of previous trials—and ultimately may provide the kind of clinical distinctions previously reserved for only the most experienced clinicians but now available to all through sensible analysis of carefully curated clinical data.

Acknowledgments

R.B.A. is supported by GM102365, LM05652, and GM61374. E.A.A. is supported by NIH OD004613 and HL105993.

Disclosures

None.

References


Key Words: Editorials ■ artificial intelligence ■ cluster analysis ■ data collection ■ informatics