VIEWPOINT

In the Era of Precision Medicine and Big Data, Who Is Normal?

Arjun K. Manrai, PhD
Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts.

Chirag J. Patel, PhD
Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts.

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Stanford University, Stanford, California; and Department of Medicine, Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California.

Supplemental content

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, Stanford University, Medical School Office Bldg, 1265 Welch Rd, Stanford, CA 94305 (jioannid@stanford.edu).

The definition of "normal" values for common laboratory tests often governs the diagnosis, treatment, and overall management of tested individuals. Some test results may depend on demographic traits of the tested population including age, race, and sex. Ideally, laboratory test results should be interpreted in reference to a population of "similar" "healthy" individuals. In many settings, however, it is unclear exactly who these individuals are. How much population stratification and what criteria for healthy individuals are optimal? In particular, with the evolution of medicine into fully personalized or "precision" medicine and the availability of large-scale data sets, there may be interest in trying to match each person to an increasingly granular normal reference population. Is this precision feasible to obtain in reliable ways and will it improve practice?

There are limited systematic analyses of baseline variation across demographically diverse population strata (including race/ancestry, gender/sex, age, and socioeconomic strata of the population) for even widespread clinical laboratory tests. Even after decades of routine use, it may be that reference standards should be reconsidered for some populations. For example, hemoglobin A_{1c} (HbA_{1c})[1] was recently found to systematically underestimate past glycemia in African American patients with the sickle cell trait.[2] There is even less documentation of whether and how more granular stratification correlates with clinical outcomes. Answering these questions would require studies that assess the outcomes of individuals with laboratory measurements classified as normal with one system vs abnormal with another. Outcomes could include both natural history and treatment benefits and harms. With limited data, small laboratory studies, and incomplete capture of long-term outcomes, this has been difficult to achieve.

> However, with the proliferation of large data sets emblematic of precision medicine, it is becoming feasible to study stratified variation and clinical outcomes at scale.

However, with the proliferation of large data sets emblematic of precision medicine,[3] it is becoming feasible to study stratified variation and clinical outcomes at scale. Sample size limitations are no longer a challenge. However, the task of defining a "normal" population becomes even more challenging. Who should define normality and using which criteria? When should standardized efforts be used across populations and instrumentation? How can multiplicity across myriad population strata be overcome as the normal population becomes more precise and personalized?

It is essential to answer these questions for widely used clinical laboratory tests such as complete blood cell count and blood chemistries before delving into more rare tests. Such tests are a routine entry point for invasive and expensive follow-up tests and procedures, yet remain poorly characterized across strata. Data sets sufficiently capacious to study stratified variation at scale include select research cohorts, electronic health records, and insurance claims data sets. Although some data sets may be queried with relative ease (eg, electronic health records at an investigator's institution or public claims data), how generalizable findings are to other clinical settings is unclear.[4]

Challenges of Precision Medicine and Big Data

Defining Normality

The first challenge to ensuring precise application of clinical laboratory testing is defining a "healthy" population to estimate the normal range of variation across population strata. A set of criteria for normality (eg, absence of chronic disease) may appear reasonable but substantial differences can result from 2 sets of equally reasonable criteria. More specifically, the Clinical and Laboratory Standards Institute (CLSI) guidelines state that 120 "reference individuals" should be used to establish reference intervals for laboratory analytes.[5] In practice, researchers and testing laboratories may use fewer than 120 individuals, often justified as sufficient to verify, rather than establish, an existing reference range. Anecdotal reports from some laboratories of major hospitals suggest that only 20 individuals may often be used for this purpose.

Furthermore, as the guideline states, health "is a relative condition lacking a universal definition." The way in which healthy individuals are defined is not standardized and the characteristics of the tested population may vary considerably between laboratories. To illustrate the potential effect of this, the US Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey (NHANES)[6] 2013-2014 survey data were examined using 3 competing definitions of normality: (1) based on the absence of common disease conditions (eg, diabetes, coronary heart disease, cancer) (62% of the NHANES population sample); (2) based on an overall excellent self-rating of health (16% of the population sample); and (3) including only individuals aged 18 to 40 years (35% of the population sample).

The 3 definitions are all defensible but lead to significant variation in the inferred normal range of HbA_{1c}, defined as lower than the 95th percentile (eFigure A in the Supplement). For example, 12%, 16%, and 27%, respectively, of all individuals would be flagged as "abnormal" using the 3 methods of defining reference ranges, based on being out of the reference range of at least 1 demographic stratum. Furthermore, using very stringent definitions for normality can lead to the paradox of "normal" becoming a rarity. For example, only 5% of the NHANES population sample have none of these disease conditions, self-rate their health as excellent, and are aged 18 to 40 years.

Multiplicity

Multiplicity across population strata causes further problems. When the distribution of an analyte (such as HbA_{1c}) is examined over many subpopulations, even when there is no difference across the subpopulations, many differences are likely to be detected if there is not a correction for the number of comparisons performed. Dealing with multiplicity is standard in some research communities, such as human genetics, but the issue is equally important in laboratory analyte comparisons in which setting reference thresholds may be much less coordinated. It is especially risky when only specific comparisons of strata are published (eg, based on having achieved statistical significance). The extent of such selective reporting biases in information on reference ranges is unknown.

A simple simulation illustrates how daunting this problem might be. Assuming there are no differences in the true analyte distributions across 5 race × 10 age × 2 gender × 3 socioeconomic = 300 population strata, if 120 individuals are repeatedly sampled (eg, from the same subpopulation) the phantom appearance of statistically significant differences will almost always be produced (many of which might seem to also have clinical relevance) even when none exist (eFigure B in the Supplement). The problem is exacerbated if the reference intervals are derived from fewer than the standard 120 individuals. The risk of erroneous inferences about reference ranges is multiplied by the number of analytes that could be tested.

Potential Solutions

The challenges involved in computing reference intervals while overcoming multiplicity can lead to suboptimal use of a test across a broad and diverse population, reducing both sensitivity and specificity and, eventually, clinical utility. Fortunately, the same large-scale data sets that present challenges to computing reference intervals (eg, electronic health records, insurance claims data) may also contain solutions. First, if longitudinal outcomes data can be reliably linked at the individual level, the clinical importance of differences in reference intervals may be testable. Second, shared large-scale databases may enable systematic analyses across data sets and laboratories while explicitly accounting for the scale of multiple testing. Third, definitions of "normal" ranges can be tailored based on patient attributes and delivered to physicians at the point of care. Fourth, and perhaps most emblematic of the precision medicine movement, computationally derived genetic ancestry (now routine to determine with genotyping arrays or sequencing) paired with laboratory testing data should allow moving beyond the often "administratively assigned" and problematic conflation of race and ancestry ubiquitous in health care data.[7]

Achieving these goals will likely benefit from the efforts of multiple groups including researchers, laboratories, health care institutions, journals, and funders. Researchers and laboratories can start by broadly sharing estimated reference intervals across demographic strata and documenting design choices such as outlier procedures and inclusion criteria, allowing other researchers to reproduce their calculations. Health care institutions could make consented patient data available to compute reference ranges for their populations. Tailored reference ranges may be possible to provide at the point of care. Journals and funders could enforce (eg, as precondition for publication or funding) or incentivize requirements that promote data sharing and explicit descriptions of selection criteria and analytic methods. Testing laboratories could share consented records that enable researchers to reevaluate claims about clinical utility across population groups. As several countries around the world embark on establishing large-scale research biobanks, it will be crucial to compute precise reference ranges and rigorously test when and how this level of precision improves care.

ARTICLE INFORMATION

Published Online: April 23, 2018. doi:10.1001/jama.2018.2009

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. International Expert Committee. International Expert Committee report on the role of the A_{1c} assay in the diagnosis of diabetes. *Diabetes Care*. 2009;32(7):1327-1334.

2. Lacy ME, Wellenius GA, Sumner AE, et al. Association of sickle cell trait with hemoglobin A_{1c} in African Americans. *JAMA*. 2017;317(5):507-515.

3. Toward NRC. *Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: National Academies Press; 2011.

4. Ioannidis JPA. Are mortality differences detected by administrative data reliable and actionable? *JAMA*. 2013;309(13):1410-1411.

5. Clinical and Laboratory Standards Institute. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline. https://clsi.org/standards/products/method-evaluation/documents/ep28/. Accessed April 6, 2018.

6. Centers for Disease Control and Prevention. National Center for Health Statistics: National Health and Nutrition Examination Survey data. https://www.cdc.gov/nchs/nhanes/. Accessed April 17, 2014.

7. Rotimi CN, Jorde LB. Ancestry and disease in the age of genomic medicine. *N Engl J Med*. 2010;363(16):1551-1558.