Commentary

From Smallpox to Big Data: The Next 100 Years of Epidemiologic Methods

Stephen J. Gange* and Elizabeth T. Golub

* Correspondence to Dr. Stephen J. Gange, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Suite W1513, Baltimore, MD 21205 (e-mail: sgange@jhu.edu).

Initially submitted May 4, 2015; accepted for publication June 3, 2015.

For more than a century, epidemiology has seen major shifts in both focus and methodology. Taking into consideration the explosion of “big data,” the advent of more sophisticated data collection and analytical tools, and the increased interest in evidence-based solutions, we present a framework that summarizes 3 fundamental domains of epidemiologic methods that are relevant for the understanding of both historical contributions and future directions in public health. First, the manner in which populations and their follow-up are defined is expanding, with greater interest in online populations whose definition does not fit the usual classification by person, place, and time. Second, traditional data collection methods, such as population-based surveillance and individual interviews, have been supplemented with advances in measurement. From biomarkers to mobile health, innovations in the measurement of exposures and diseases enable refined accuracy of data collection. Lastly, the comparison of populations is at the heart of epidemiologic methodology. Risk factor epidemiology, prediction methods, and causal inference strategies are areas in which the field is continuing to make significant contributions to public health.

The framework presented herein articulates the multifaceted ways in which epidemiologic methods make such contributions and can continue to do so as we embark upon the next 100 years.

inference; measurement; populations; synthesis

For more than a century, epidemiology has seen major shifts in both focus (1) and methodology (2). The roots of epidemiology focused on mortality outcomes and clusters of symptomatic cases in the sanitation era of the 1800s; the creation of germ theory set the stage for expansion, allowing identification of determinants of specific infectious diseases, such as smallpox and cholera. This, in turn, served as the foundation for mid-20th century interest in the epidemiology of chronic diseases, other health outcomes such as trauma (3), and the myriad of factors that contribute to health and disease. In the modern era, the most recent edition of the Dictionary of Epidemiology has acknowledged the breadth of topics to which epidemiology now applies by defining epidemiology as the study of “. . . health-related events, states, and processes [that] include outbreaks, diseases, disorders, causes of death, behaviors, environmental and socioeconomic processes, effects of preventive programs, and use of health and social services” (4, p. 95).

Along with these evolving trends in application, epidemiologic methods have also expanded substantially. Classical approaches for identifying determinants of disease that built upon biological pathogenesis were augmented with statistically based approaches in which “black-box” understanding demonstrated both value and deficiencies (5). Simple characterization of outcomes, such as Snow’s classic map of cholera cases in London, have been replaced with modern surveillance systems and other health information systems (6). Modern epidemiology has embraced the growing set of causal inference tools (7) and expanded frameworks beyond the simple “epidemiological triad” for use in understanding disease processes, as illustrated by the life-course (8) and multilevel eco-epidemiology (1) models.

Throughout this evolution in focus and methodology, what have been the common tenets of epidemiologic methods and how will epidemiology change over the next 100 years? Morabia (2) points to 2 common features of epidemiologic methods: population thinking and group comparisons. Building upon this foundation, we have promulgated a framework for epidemiology methods (9) incorporating 3 domains:

- Populations, study design, and follow-up;
- Measurement; and
- Comparison of populations and inference
Figure 1 emphasizes the interconnection between these 3 components; this, along with synthesis, serves as the basis for understanding and action in public health. We believe this framework is valuable not only for describing the essence of epidemiologic methods currently but also for conceptualizing how epidemiology will remain relevant in the future. In this article, we discuss each component of the framework in the context of emerging trends in epidemiologic methodology, particularly in light of increased interest in the promise (and pitfalls) of “big data.”

POPULATIONS, STUDY DESIGN, AND FOLLOW-UP

Population thinking in epidemiology, as in other population-based disciplines like demography and sociology, distinguishes the field from those like medicine, in which the focus is on the individual. However, the kind of population that is amenable for epidemiologic inquiry is entirely unconstrained. Public health agencies are charged with using approaches (17), and patient-oriented outcomes (18), innovative experimental and observational designs are needed to generate the data relevant for evidence-based decisions. Epidemiology has a strong role to play in these new designs, with a key strength being an explicit acknowledgement that populations are not static but rather change through time (19). By framing study designs in the context of time, an epidemiologic perspective equips one to determine the influence of different design features (such as eligibility, follow-up, timing of measurements, etc.) more than an informatics perspective (which often focuses on the logistics of data capture) or a perspective focused on data analysis does.

MEASUREMENT

In addition to greater availability of data on populations, innovations in which data are measured have spurred advances in epidemiologic methods. Measurements of exposures, outcomes, potential confounders, and other variables of interest are germane to epidemiology; hence, they comprise a major domain of epidemiologic methods in our framework, and that domain is the one in which we have seen great changes over the past century. Because inaccurate measurement can be a significant source of bias, issues related to measurement methods and assessment of measurement validity are major foci of constant refinement in epidemiologic methods. Although we strive for accuracy, reality often dictates an inverse correlation between the accuracy and practicality of data collection (20). Thus, the importance of identifying appropriate measurement tools cannot be understated. Moreover, as advances in technology enable novel approaches to measurement, it becomes increasingly important to evaluate the validity of these new metrics.

Traditional data collection methods such as population-based surveillance and individual interview have been supplemented over the past century with the development of new methodologies. The ability to measure biological markers of exposures and diseases at increasingly finer levels has greatly
expanded our understanding of disease etiology. Identified biomarkers can function as measures of environmental exposure (21), measures of vaccination history (22), markers of psychiatric illness (23) or markers of disease progression (24, 25). Biologic measures of exposures and disease have evolved to include genomics, metabolomics, and epigenetics. Genome-wide association studies have brought the promise of evaluating the entire genome for regions that might influence the risk of a particular disease.

Over time, there have also been innovations in measures other than biomarkers. Whereas investigation of single exposures via personal interview once gave way to health systems databases and electronic medical records, current innovations enable measurement of the environment and constellations of exposures. Environmental sensors allow for assessment of neighborhood-level effects on health. Most recently, the mobile health movement (also referred to as mHealth) has led to new technologies for health measurements, such as wearable devices (e.g., heart monitors, wearable pedometers, activity trackers) and mobile phone applications that allow users to track their diet, sleep, weight, and exercise (26), that have resulted in individual members of study populations doing the measurement for us.

The importance of accuracy in measurement and the rate at which measurement tools are evolving were underscored by the 2010 call for a new type of scholarly paper on “methods of measurement in epidemiology” by the International Journal of Epidemiology (27). In the future, this area will continue to evolve, starting with scaling programs that use mHealth in innovative ways, particularly in low- and middle-income countries (28) and among hard-to-reach populations (29).

COMPARISON OF POPULATIONS AND INFERENCE

Similar to Morabia’s foundation for epidemiologic methods (2), our framework also highlights population comparisons as a central domain of epidemiologic methods. Epidemiologic training conditions us to always ask “compared with whom?” when data for a group are presented. The use of good epidiologic methods is crucial when identifying appropriate comparison groups, especially when the task is difficult (30).

Central to these comparisons are the foundations of causal inference, which has matured over the past several decades to stand as a distinct discipline. Epidemiology has an interest in causal inference methods (as do many other disciplines, such as statistics, economics, sociology, and education), and it has also been a fertile field in which significant contributions to the methods have been made. The development and adaptation of formal definitions, frameworks, and modeling strategies has clearly benefited epidemiology and will likely continue to do so in the future. Causal inference has only begun to catch up to the emergence of big data (31), and this will likely be a fruitful area for future research and application.

Despite the importance of causal inference to epidemiology, the field has long embraced other approaches that are not steeped in a causal framework. “Risk-factor epidemiology” approaches have long utilized statistical tools in an attempt to identify all of the important predictors of an outcome. These methods continue to be used to reveal associations, particularly when there is little understanding of underlying mechanisms or models (5). These inferred associations are then used as the basis for constructing theory and/or causal models. The increased availability of software for conducting these analyses, including those that implement sophisticated machine-learning algorithms (32) and model-building strategies (33), makes it likely that these approaches will continue to be part of the fabric of epidemiologic inference. The difficulties with using these approaches to form a cohesive picture will likely be magnified in the era of big data, in which there is much variability in the quantity and quality of data, as illustrated in the search for independent risk factors by Brotman et al. (34).

Another area in inferential statistics in which epidemiology has played an important role is in the development of prediction models that utilize statistical approaches not as the basis for theoretical or causal inference but to maximize the ability to predict outcomes for patients at the individual level. Although many prediction models have been developed, it is likely the number of models will swell with the combination of big data and heightened interest in precision medicine (35). Epidemiologists should be at the forefront of the push to use appropriate methodology for assessing the accuracy, precision, and effectiveness of these inferences.

CONCLUSIONS

There have long been calls for self-examination to ensure that our discipline’s methods and frameworks are up to the challenge of making a difference in the health of society (36–38). With the explosion of big data, the development of more sophisticated data collection and analytical tools, and the increased pressure to use these data for evidence-based solutions, these periodic assessments seem more important than ever. We believe the framework that we have outlined above provides a summary of fundamental domains that is simplistic yet articulates how epidemiologic methods have contributed to date and will continue to contribute in the future (39). The framework also provides a guide for how to organize the pedagogy of our introduction of epidemiologic methods in a balanced manner that is dominated by neither any one particular domain nor any particular methodology (40). By inculcating these domains in our teaching, we provide students a framework that allows them to adapt with changing times and innovations, from understanding and eradicating “small pox” to the current challenges to health in the era of “big data.”

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland (Stephen J. Gange, Elizabeth T. Golub).

This work was supported by the Johns Hopkins University Gateway Sciences Initiative.

Conflict of interest: none declared.

REFERENCES


