

# Surveillance as Our Sextant

*AJPH* is inaugurating a new section on public health surveillance and survey methods, no doubt in part to showcase the different ways surveillance is breaking methodological ground and advancing frontiers of public health knowledge with the recent explosive growth in data availability and computational tools. The not-implausible vision for tomorrow's population health surveillance is one in which jurisdictions large and small can harness a lattice of highly sensitive information systems that provide near-real-time information on wide ranges of health conditions—from novel influenza detection to politically induced stress—to guide health programs, policies, and research.

Yet, important challenges in the field of population health surveillance still remain despite the tantalizing array of new data streams and technologies. Efforts to address stubborn time lags, inaccuracies, oppressive costs, and data incompleteness are as relevant today as they were 20 to 50 years ago. For both exciting innovations and persistent limitations, the new *AJPH* section is a greatly welcome addition. The science of advancing designs and approaches in monitoring disease requires dialogue, peer scrutiny, and replication. As a population health surveillance methodologist of more than 20 years, I would flag two important priorities to catalyze population health surveillance functionality: (1) realizing the promise of big data from health care and (2) improving small area estimation.

## REALIZING PROMISES OF BIG HEALTH CARE DATA

One great hope for surveillance is to harness big data generated routinely by today's health care system, including electronic health records (EHRs) and claims databases. The speed at which this field is evolving is both lightning fast—particularly with respect to infrastructure and computing capacity—and glacially slow. The drag is mostly attributable to the lack of governance models, insufficient funding, and a dearth of validation studies to guide system improvement.

On the rapid advancement front, the infrastructure for tapping big health care data has been spurred by dramatic EHR uptake since 2004, financial incentives to achieve meaningful-use reporting standards, and federally funded initiatives to build nationwide data networks between health care systems. These, together with high-performance computing tools such as automated rapid-cycle queries, natural language processing, and machine learning, have stimulated exciting new surveillance initiatives.

Take the US Food and Drug Administration's (FDA's) Sentinel Initiative, for example. Sentinel is an active drug safety monitoring surveillance system that provides the FDA with near-real-time access to claims data on more than 200 million patients and 5.5 billion patient encounters across the country.<sup>1</sup> The vast size and speed of the system allows the FDA to quickly identify drug or device safety concerns. The FDA contracts with an academic partner to collate responses from 18 major health

insurers working from a standardized data model, bypassing need for competitors to share protected information.

A second innovative big data surveillance system is the Durkheim Project,<sup>2</sup> designed to tap into Veterans Affairs medical records and opt-in social media by using natural language process and machine learning techniques. The goal is to allow clinicians to evaluate risk of suicide during patient encounters through predictive models.

A third innovative surveillance example is the NYC Macroscopic, a municipal surveillance system designed to query primary care practices in a large distributed EHR network,<sup>3</sup> which is described by Perlman et al. in this issue of *AJPH* (p. 853). Practice-level count information on common chronic conditions is tabulated and statistically weighted to represent the sociodemographic profile of New York City adults, and prevalence estimates of conditions such as hypertension, diabetes, high cholesterol, obesity, and smoking are generated at lower cost, some of which rival survey methods in accuracy. These and other examples demonstrate the potential of large-scale health care data for surveillance while still accommodating privacy considerations by using aggregated queries at the practice or payer level or modeling.

Yet, despite tremendous innovations, each of these systems is still undergoing testing, and tangible yields for public health remain limited. Although the Sentinel Initiative was first piloted in 2008, the system still serves as a secondary data source to traditional provider reporting methods, and queries have only produced a handful of minor FDA actions.<sup>4</sup> Part of the delay in uptake has been insufficient validation and refinement to identify what aspects are accurate and reliable for prime-time use. It would be exciting to see such studies presented in these pages. The Durkheim Project, despite promising results in its demonstration project, is stalled from lack of funding. NYC Macroscopic is operational after 5 years of planning and validation, yet additional work is needed to determine its accuracy in analyzing trends over time, especially in the face of rapidly changing EHR use. Overall, however, the pace of progress in harnessing big data for surveillance is on a decidedly upward trajectory, and *AJPH* is poised to accelerate that rapid learning by widening the audience.

## IMPROVING SMALL AREA ESTIMATION

Small area estimation is currently one of the most exciting growth areas in surveillance methods. We frequently obtain critical information on health

### ABOUT THE AUTHOR

Lorna E. Thorpe directs the Division of Epidemiology at the New York University School of Medicine, Department of Population Health, New York, NY.

Correspondence should be sent to Lorna E. Thorpe, Professor, NYU School of Medicine Division of Epidemiology, 650 1st Ave, Rm 512, New York, NY 10016 (e-mail: lorna.thorpe@nyumc.org). Reprints can be ordered at <http://www.ajph.org> by clicking the "Reprints" link.

This editorial was accepted March 18, 2017.  
doi: 10.2105/AJPH.2017.303803

behaviors and conditions from carefully designed population-based surveys. These surveys remain a bedrock in the arsenal of chronic disease surveillance, yet, because of high costs, most survey-based surveillance sources are insufficiently granular for detailed analysis at the geographic level (e.g., neighborhood, city) where policy and programmatic action may be most feasible and have greatest impact. Small area estimation typically involves combining data from multiple sources such as surveys, census, and administrative records, sometimes using statistical models to estimate local-area outcomes with adequate precision by “borrowing strength” from the other sources. Results are then synthetic, but they can provide new, important information to local policymakers who previously lacked data.

A recent large-scale example of such an effort is the 500 Cities Project by the Centers for Disease Control and Prevention and Robert Wood Johnson Foundation.<sup>5</sup> By combining the Centers for Disease Control and Prevention’s Behavioral Risk Factor Surveillance System results with US Census and American

Community Survey data in a multilevel logistic regression model and applying demographic poststratification weights at the census tract level, the project models city- and census tract-level estimates for 27 chronic disease measures for the 500 largest American cities.

Keeping pace with these innovations in small area estimation is the growing list of problems to be resolved. Perhaps most pressing is that small area estimation is still time-intensive because of complex model specification demands and software limitations with respect to run-time and convergence. As a consequence, only a few initiatives currently provide annual estimates by using validated, cutting-edge methods. A second problem is the need to compare small area estimation approaches and evaluate inherent biases. For example, studies indicate that small area estimation models perform better in areas where population characteristics align with “average” distributions and are less accurate in “atypical” areas, such as large, diverse cities.<sup>6,7</sup>

Finally, although modeled estimates are an important new resource, local practitioners ultimately desire actual estimates to allocate

resources and evaluate programs. To accompany advances in estimation, the broader public health community should support efforts to sustain and expand access to federal microdata through public use data sets, and promote approaches to analyze underlying microdata while maintaining confidentiality, such as through user-friendly, interactive electronic dashboards.

The ultimate goal of population health surveillance is to gain actionable insight. As we face widening health inequalities, intransient disease management improvements in health care quality, and threats of new pandemics, we need accurate, reliable, and timely data. I applaud *AJPH* for embracing the need to focus on surveillance methodology. Far from being viewed as a simple and unexciting function, surveillance is our sextant, without which we are unmoored. *AJPH*

Lorna E. Thorpe, PhD

#### ACKNOWLEDGMENTS

L. E. Thorpe’s efforts are supported in part by the Centers for Disease Control and Prevention (grant U48DP001904).

Thank you to Marc Gourevitch, Jessica Athens, and Alexis Feinberg for their

review and comment on an early draft of this article.

#### REFERENCES

- Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System—a national resource for evidence development. *N Engl J Med*. 2011;364(6):498–499.
- Poulin C, Thompson P, Bryan C. Public health surveillance: predictive analytics and big data. In: *Artificial Intelligence in Behavioral and Mental Health Care*. San Diego, CA: Academic Press; 2016: 205–230.
- Newton-Dame R, McVeigh KH, Schreiberstein L, et al. Design of the New York City MacroScope: innovations in population health surveillance using electronic health records. *EGEMS (Wash DC)*. 2016;4(1):1265.
- Kuehn BM. FDA’s foray into big data still maturing. *JAMA*. 2016;315(18): 1934–1936.
- National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. 500 Cities: local data for better health. Centers for Disease Control and Prevention. 2016. Available at: <https://www.cdc.gov/500cities>. Accessed March 22, 2017.
- Chin S-F, Harding A. *Regional Dimensions: Creating Synthetic Small-Area Microdata and Spatial Microsimulation Models*. Canberra, Australia: National Centre for Social and Economic Modelling; 2006.
- Smith DM, Pearce JR, Harland K. Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health Place*. 2011;17(2):618–624.

Applied Physics Laboratory, as consultant to the US Centers for Disease Control and Prevention, and as board member and research committee chair of the International Society for Disease Surveillance. These efforts have been enriched by collaborations with US and international public health partners, both civilian and military.

## Evolution of Public Health Surveillance: Status and Recommendations

Public health institutions at local, regional, and national levels face evolving challenges with limited resources. Multiple forms of data are increasingly available, ranging from streaming statistical data to episodic reports of confirmed disease incidence. While technological tools for collecting and using these data proliferate, economic pressures often

preclude growth of concomitant staff with required expertise. The intent here is to provide perspective on evolution of public health surveillance since the late 1990s, suggest how methodological approaches can be improved, and recommend areas of growth given mandates for evidence-based policy and practice.

Remarks in this article stem from my last 18 years’ work on surveillance system development at the Johns Hopkins University

#### ABOUT THE AUTHOR

Howard S. Burkom is with the Research and Exploratory Development Department at The Johns Hopkins Applied Physics Laboratory, Laurel, MD.

Correspondence should be sent to Howard S. Burkom, Johns Hopkins Applied Physics Laboratory, 11100 Johns Hopkins Rd, Laurel, MD 20723 (e-mail: [howard.burkom@jhuapl.edu](mailto:howard.burkom@jhuapl.edu)). Reprints can be ordered at <http://www.ajph.org> by clicking the “Reprints” link.

This editorial was accepted March 16, 2017.

doi: 10.2105/AJPH.2017.303801