

# State and Local Chronic Disease Surveillance Using Electronic Health Record Systems

Michael Klompas, MD, MPH, Noelle M. Cocoros, DSc, MPH, John T. Menchaca, BS, Diana Erani, MBA, Ellen Hafer, MTS, MBA, Brian Herrick, MD, Mark Josephson, BS, Michael Lee, MD, Michelle D. Payne Weiss, MPH, Bob Zambarano, PhD, Karen R. Eberhardt, BS, Jessica Malenfant, MPH, Laura Nasuti, PhD, MPH, and Thomas Land, PhD

**Objectives.** To assess the feasibility of chronic disease surveillance using distributed analysis of electronic health records and to compare results with Behavioral Risk Factor Surveillance System (BRFSS) state and small-area estimates.

**Methods.** We queried the electronic health records of 3 independent Massachusetts-based practice groups using a distributed analysis tool called MDPHnet to measure the prevalence of diabetes, asthma, smoking, hypertension, and obesity in adults for the state and 13 cities. We adjusted observed rates for age, gender, and race/ethnicity relative to census data and compared them with BRFSS state and small-area estimates.

**Results.** The MDPHnet population under surveillance included 1 073 545 adults (21.8% of the state adult population). MDPHnet and BRFSS state-level estimates were similar: 9.4% versus 9.7% for diabetes, 10.0% versus 12.0% for asthma, 13.5% versus 14.7% for smoking, 26.3% versus 29.6% for hypertension, and 22.8% versus 23.8% for obesity. Correlation coefficients for MDPHnet versus BRFSS small-area estimates ranged from 0.890 for diabetes to 0.646 for obesity.

**Conclusions.** Chronic disease surveillance using electronic health record data is feasible and generates estimates comparable with BRFSS state and small-area estimates. (*Am J Public Health*. Published online ahead of print July 20, 2017: e1–e7. doi:10.2105/AJPH.2017.303874)

**P**ublic health agencies, clinicians, regulators, and community organizations require timely and accurate data on the prevalence and incidence of key health indicators at the state and local level to identify health disparities, trigger interventions, and evaluate programs.<sup>1</sup> Officials have traditionally used death records, claims databases, and specialized surveys to monitor chronic diseases. Two surveys in particular account for the majority of US public health agencies' data on chronic conditions: the Behavioral Risk Factor Surveillance System (BRFSS) and the National Health and Nutrition Examination Survey (NHANES).<sup>2–4</sup> Both of these surveys provide rich data on health conditions and health behaviors but also have important limitations in their timeliness, accuracy, breadth, and cost.

BRFSS is a national continuous random-digit-dial landline and cellphone survey that collects self-reported data on health outcomes and health behaviors from approximately

500 000 adults each year.<sup>3,5</sup> Strengths of the system include its large sample size and the inclusion of data on health-related behaviors that are poorly documented in paper and electronic medical records (e.g., food choices, exercise, seatbelt use). All states and many large cities participate in the survey. The limitations of BRFSS include a 2- to 3-year delay between data collection and data publication, lack of data on children and people without telephones, reliance on self-reports rather than clinical measures, limited

capacity to estimate the prevalence of rare conditions, limited coverage of smaller cities and towns, and limited coverage of smaller demographic groups such as Native Americans.<sup>3,5–7</sup>

NHANES, by contrast, organizes in-person testing for a nationally representative sample of adults.<sup>8,9</sup> The evaluation includes both a questionnaire and detailed medical, dental, and physiological measurements. The sample includes about 5000 people per year. NHANES provides very rich clinical and behavioral data, but the survey is expensive and its small sample size precludes the possibility of estimating state and local disease rates, tracking rare conditions, or estimating disease rates in some minority populations. As with BRFSS, there is typically an interval of years between data acquisition and data publication.

Some states have used small-area estimation techniques to derive more granular estimates at the county, community, or neighborhood level using BRFSS data combined with US Census data. Recent funding cuts for BRFSS, however, have decreased some states' sample sizes to the point that small-area estimation is no longer feasible without combining multiple surveys across multiple years.

The increasingly widespread adoption of electronic health record (EHR) systems provides a potential new surveillance option

## ABOUT THE AUTHORS

Michael Klompas, Noelle M. Cocoros, John T. Menchaca, and Jessica Malenfant are with the Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA. Diana Erani, Ellen Hafer, and Mark Josephson are with the Massachusetts League of Community Health Centers, Boston. Brian Herrick and Michelle D. Payne Weiss are with Cambridge Health Alliance, Cambridge, MA. Michael Lee is with Atrius Health, Boston. Bob Zambarano and Karen R. Eberhardt are with Commonwealth Informatics Inc, Waltham, MA. Laura Nasuti and Thomas Land are with the Office of Data Management and Outcomes Assessment, Massachusetts Department of Public Health, Boston.

Correspondence should be sent to Michael Klompas, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Dr, Suite 401 East, Boston, MA 02215 (e-mail: mklompas@partners.org). Reprints can be ordered at <http://www.ajph.org> by clicking the "Reprints" link.

This article was accepted April 23, 2017.

doi: 10.2105/AJPH.2017.303874

to complement BRFSS and NHANES.<sup>10–15</sup> EHRs include detailed clinical data on very large populations and are updated continually. In addition, the data are in electronic format and arrayed in structured fields, so it is possible to create programs that can automatically analyze the data on a routine basis to produce timely, granular, and detailed surveillance summaries.<sup>14</sup>

A number of model EHR-based surveillance systems now exist, but little is known about whether they can provide reliable estimates of disease prevalence at the state or local level.<sup>16</sup> A major theoretical limitation of EHRs is that population coverage is not random. EHRs only include data on patients affiliated with a specific clinical practice (which may be biased by factors such as location, target population, services offered, insurance types accepted, and population insurance coverage rates).<sup>17</sup> In addition, EHRs include only data on the subset of the population that seeks care, and on the tests and conditions that clinicians order and diagnose in the course of routine care (rather than systematically testing all patients). These factors may bias EHRs toward greater coverage of women, children, and sicker individuals.

Given these questions, we sought to characterize the capacity of EHRs to estimate state and local disease rates in Massachusetts. We analyzed 5 test conditions—diabetes, asthma, smoking, hypertension, and obesity—using EHR-data derived from the MDPHnet public health surveillance platform and compared results with BRFSS state and small-area estimates using the Centers for Disease Control and Prevention's (CDC's) 500 Cities data for 13 Massachusetts communities.<sup>10,11</sup>

## METHODS

MDPHnet is a distributed health data network in Massachusetts that permits authorized public health officials to submit detailed queries against the EHR data of 3 large multispecialty group practices that serve a combined patient population of approximately 1.5 million people (about 20% of the state population). Each practice uses open-source software (electronic medical record Support for Public Health; <http://www.esphhealth.org>) to create, host, and maintain

a data repository behind its firewall in accordance with a common data specification.<sup>10,18</sup> MDPHnet then uses a software package called PopMedNet (<http://www.PopMedNet.org>) to securely distribute queries to each clinical partner to run locally against its repository.<sup>19,20</sup> Public health officials enter their queries into a Web-based portal that then sends the queries to each partner. Clinical partners have the option to review queries before permitting their execution and to review results before allowing them to be released to the health department. Partners can also permit queries to execute automatically. Results are returned to the health department as aggregate counts without personal identifiers.

MDPHnet's current practice partners include a large multisite, multispecialty ambulatory group serving primarily a well-insured population (approximately 800 000 patients, 29 clinical locations), a network of community health centers targeting underserved populations (approximately 500 000 patients, 15 clinical locations), and a combination safety net—general practice hospital and ambulatory group (approximately 200 000 patients, 15 clinical locations). Partner practices use a variety of different EHR systems. The overall patient return rate across all 3 practice groups is 82%.

To estimate disease prevalence, MDPHnet users submit 2 queries: (1) to find the count of patients with a specified outcome (observed patients, the numerator) and (2) to determine the total count of patients in the population (observed population, the denominator). MDPHnet users can define diseases of interest (the numerator) using combinations of vital signs, laboratory test results, diagnosis codes, and prescriptions. The observed population (the denominator) is defined as the count of patients with at least 1 outpatient visit of any kind in the health care system in the 2 years prior to the index date of the query.

MDPHnet users can adjust disease prevalence estimates for age (10-year bands), gender, and race/ethnicity by having MDPHnet stratify the observed population by these factors. Within each stratum, MDPHnet projects the number of patients with the disease of interest by multiplying the prevalence of the disease in each stratum per MDPHnet by the total number of people in the stratum per the 2010 US census. Users can

then calculate the projected number of patients with the disease across all strata by summing the projected counts per stratum. The adjusted prevalence of disease is the projected number of patients with the disease of interest divided by the size of the census population. Users can limit queries to select regions or towns in Massachusetts to generate observed and adjusted estimates for a particular location. Patients are assigned to a location on the basis of their home zip code. Patients' ages are calculated as of the index date of the query.

We assessed the feasibility and comparability of MDPHnet disease estimates using diabetes (type 1 and type 2 combined), asthma, smoking, hypertension, and obesity in adults as test cases. We defined diseases using combinations of diagnosis codes, vital signs, laboratory tests, and prescriptions as appropriate to each condition (see the box on page e3). We designed these algorithms through consultation with expert clinicians and public health practitioners and validated them through targeted chart reviews.<sup>21</sup> We submitted queries for the 5 test conditions via MDPHnet to the 3 participating clinical practices, using January 1, 2014, as the index date for the query. We limited all queries to patients with at least 1 encounter in the 2 years preceding the index date to focus the evaluation as well as possible on patients actively associated with each practice. We chose the index date of the query (January 1, 2014) to facilitate comparison with BRFSS state and small-area estimates (MDPHnet data are updated weekly, but 2014 is the most recent year for which BRFSS data are available). We adjusted MDPHnet surveillance results for age, gender, and race/ethnicity. We aggregated results from each clinical partner for analysis.

We compared MDPHnet adjusted prevalence estimates for each of the 5 test conditions against statewide BRFSS data and small-area estimates. We obtained small-area estimates from CDC's 500 Cities initiative, which includes data on 13 Massachusetts communities.<sup>22</sup> CDC derived the 500 Cities small-area estimates by developing regression models to relate census data to BRFSS survey data at the state level and then applying the regression models to census data from small areas to predict local disease rates.<sup>23,24</sup> We compared city-level MDPHnet crude and adjusted disease prevalence estimates with the

**MDPHNET DISEASE IDENTIFICATION CRITERIA****Condition and Definition****Diabetes ( $\geq 1$  of the following):**

- Hemoglobin A1C  $\geq 6.5$
- Fasting glucose  $\geq 126$
- Random glucose  $\geq 200$  on 2 or more occasions
- Prescription for insulin outside of pregnancy
- ICD-9 code 250.xx or ICD-10 code E10\*, E11\*, or E14\* on 2 or more occasions
- Prescription for glyburide, gliclazide, glipizide, glimepiride, pioglitazone, rosiglitazone, repaglinide, nateglinide, meglitinide, sitagliptin, exenatide, alogliptin, linagliptin, saxagliptin, albiglutide, dulaglutide, liraglutide, canagliflozin, dapagliflozin, empagliflozin, or pramlintide

**Asthma ( $\geq 1$  of the following):**

- $\geq 2$  encounters with diagnosis code for asthma (ICD-9 493.xx or ICD-10 J45\*-J46\*)
- $\geq 2$  prescriptions for albuterol, levalbuterol, pirbuterol, arformoterol, formoterol, indacaterol, salmeterol, beclomethasone, inhaled budesonide, inhaled ciclesonide, inhaled flunisolide, inhaled fluticasone, inhaled mometasone, montelukast, zafirlukast, zileuton, ipratropium, tiotropium, cromolyn INH, omalizumab, fluticasone + salmeterol, albuterol + ipratropium, mometasone + formoterol, or budesonide + formoterol

**Smoking:**

- Most recent smoking status of "current" as recorded in the EHR "tobacco" field (synonyms for current acceptable)

**Hypertension ( $\geq 1$  of the following):**

- Systolic blood pressure  $\geq 140$  mm Hg or diastolic blood pressure  $\geq 90$  on  $\geq 2$  occasions within a 1-y period
- Diagnosis code for hypertension (ICD-9 401.xx or 405.xx or ICD-10 I10 or I15) and prescription for at least 1 of the following medications within 1 y of the hypertension diagnosis code: hydrochlorothiazide, indapamide, amlodipine, clevidipine, felodipine, isradipine, nicardipine, nifedipine, nisoldipine, diltiazem, verapamil, acebutolol, atenolol, betaxolol, bisoprolol, carvedilol, labetalol, metoprolol, nadolol, nebivolol, pindolol, propranolol, benazepril, captopril, enalapril, fosinopril, lisinopril, moexipril, perindopril, quinapril, ramipril,trandolapril, candesartan, eprosartan, irbesartan, losartan, olmesartan, telmisartan, valsartan, clonidine, doxazosin, guanfacine, or methyl dopa

**Obesity:**

- Most recent BMI  $\geq 30$  kg/m<sup>2</sup>
- If BMI not available, calculate using last available weight measured within the last y and last available height measured when aged  $\geq 16$  y

Note. ICD-9 (10) = International Classification of Diseases, Ninth (Tenth) Revision; EHR = electronic health record; BMI = body mass index (defined as weight in kilograms divided by the square of height in meters).

BRFSS 500 Cities small-area estimates using Pearson correlation coefficients. We assessed associations between MDPHnet's census coverage rate per city and the relative difference between MDPHnet and BRFSS small-area estimates by scatter plot.

**RESULTS**

The MDPHnet population under surveillance included 1 073 545 adults aged 20

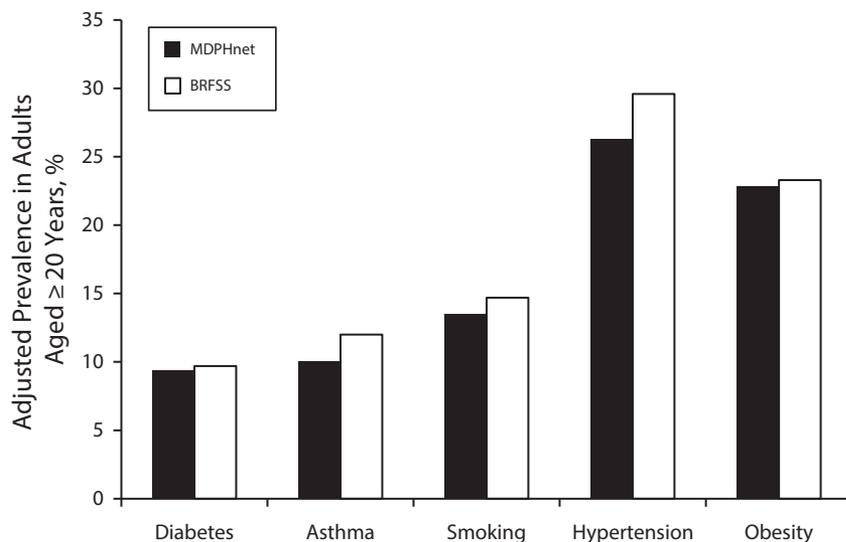
years and older with at least 1 encounter in the 2 years preceding January 1, 2014. This corresponds to 21.8% of the Massachusetts adult census population from 2010. Absolute sample sizes for MDPHnet in the 13 comparison cities included in the study ranged from 410 (Springfield) to 138 369 (Boston). The percentage of the adult census population included in MDPHnet samples ranged from 0.4% (Springfield) to 63.2% (Somerville). The median sample size for

MDPHnet among the 13 comparison cities was 25 702 and the median census coverage rate was 31.5%. By way of comparison, the Massachusetts BRFSS survey from 2014 included 15 654 adults statewide (0.3% of the Massachusetts adult population). The count of Massachusetts BRFSS respondents by city was not available.

Statewide estimates of the prevalence of diabetes, smoking, asthma, hypertension, and obesity using MDPHnet are shown in Figure 1. MDPHnet state-level estimates were similar to those in BRFSS: diabetes, 9.4% versus 9.7% in MDPHnet and BRFSS, respectively (95% confidence interval [CI] = 9.0%, 10.4%); asthma, 10.0% versus 12.0% (95% CI = 11.1%, 12.8%); smoking, 13.5% versus 14.7% (95% CI = 13.7, 15.7); hypertension, 26.3% versus 29.6% (95% CI = 28.3, 30.5); and obesity, 22.8% versus 23.3% (95% CI = 22.3%, 24.4%).

City-level estimates for the 5 target conditions in MDPHnet and CDC's 500 Cities BRFSS small-area estimates are tabulated in Table 1. In addition, Figure 2 plots the correspondence between MDPHnet estimates and BRFSS small-area estimates for all cities and all conditions. Correlation coefficients for MDPHnet city-level estimates versus BRFSS small-area estimates ranged from 0.646 to 0.890. Estimates matched best for diabetes (correlation coefficient = 0.890) and hypertension (correlation coefficient = 0.828). There was greater variability in the estimates for smoking (correlation coefficient = 0.747), asthma (correlation coefficient = 0.672), and obesity (correlation coefficient = 0.646). For smoking in particular, there were a few towns in which MDPHnet estimates were less than half the corresponding BRFSS small-area estimates.

Pearson correlation coefficients for MDPHnet crude and adjusted prevalence estimates versus BRFSS small-area estimates are shown at the bottom of Table 1. Adjusting for differences in the age, gender, and race/ethnicity distribution between the MDPHnet observed population and state and local census populations improved the correlation between MDPHnet and BRFSS small-area estimates for all conditions. There was limited correlation between MDPHnet census coverage rates and the relative differences between MDPHnet estimates and BRFSS small-area estimates (Figure A, available as



**FIGURE 1—Comparison of Selected Disease Prevalences Using MDPHnet vs Behavioral Risk Factor Surveillance System (BRFSS): Massachusetts, 2014**

a supplement to the online version of this article at <http://www.ajph.org>.

## DISCUSSION

Tracking the prevalence of key conditions and health behaviors at the state and community level is a core need for public health agencies. Health departments have traditionally relied on BRFSS and NHANES to estimate disease rates, but these instruments have many limitations, including high cost, limited sample sizes, substantial delay between data acquisition and publication, and, in the case of BRFSS, reliance on self-reports. We demonstrate that automated analyses of EHR data held by large practices compare favorably with traditional surveys. MDPHnet's state-level estimates of the prevalence of diabetes, asthma, smoking, hypertension, and obesity closely matched BRFSS estimates.

MDPHnet disease estimates at the city level showed moderate to high levels of correlation with CDC's small-area estimates, albeit with some significant outliers.

### Advantages of EHR-Based Surveillance

There are many potential advantages to using routinely collected EHR data for

routine chronic disease surveillance. The data are timelier than traditional surveys (MDPHnet data are updated weekly), they include clinical measures rather than self-reports, they include data on very large numbers of patients, and they are available in digital format, thus speeding acquisition and readiness for analysis. The timeliness of the data allows for the possibility of evaluating public health interventions in near real time to inform the need for midcourse corrections to optimize the probability of success. The inclusion of data from very large numbers of patients permits more granular analyses than are possible with traditional surveys, including small-area estimates using direct observations rather than statistical extrapolations, surveillance for rare diseases (e.g., type 1 diabetes), and monitoring the health of minority populations. In addition, EHR data can be used to evaluate patterns of care, health care utilization, and laboratory parameters that are either too complicated or too onerous to collect using manual surveys. For example, one could evaluate the percentage of patients with diabetes who have been evaluated for microalbuminuria, the percentage of patients with asthma prescribed controller medications, or the percentage of patients with hypertension under adequate control.

A major concern with using EHR data to track chronic disease rates has been the

question of whether EHR-based disease estimates are representative of the general population given that the populations covered by EHRs are not random and the data in EHRs are based on clinicians' targeted assessments of selected patients rather than total population screening.<sup>17</sup> Our analysis suggests that EHR-based estimates drawn from multiple large populations can approximate traditional surveillance sources well, particularly at the state level. Some of the factors that may contribute to the robustness of MDPHnet's state-level estimates include the very large number of patients under surveillance (approximately 20% of the state population), the inclusion of data from a diverse set of practices (including a practice geared toward privately insured middle-income patients, a safety net provider, and a network of community health centers), and the availability of tools to adjust estimates for differences in the age, gender, and race/ethnicity distribution between the observed population and the census population of the target surveillance area. The high rate of insurance coverage in Massachusetts may also account for some of the success of MDPHnet by increasing the breadth of patients engaged in clinical care.

### Sources of Error in MDPHnet and 500 Cities Estimates

Correlations between MDPHnet's city-level estimates and CDC's 500 Cities small-area estimates were moderate to high overall, but there were nonetheless some major discrepancies between these 2 sources for some diseases in some cities. Possible reasons for these discrepancies may lie with MDPHnet, 500 Cities small-area estimates, or both.

Possible sources of error within MDPHnet could include low levels of population coverage in some communities (e.g., Springfield, where MDPHnet coverage is 0.4% of the adult population), poor documentation in the electronic record (e.g., smoking status, body mass index), and reliance on data from a small number of clinical partners at the city level (e.g., some MDPHnet communities are only served by 1 MDPHnet-affiliated clinic, which may in turn serve a limited and non-representative subset of the community). MDPHnet case definitions may misclassify some patients, and MDPHnet may err in

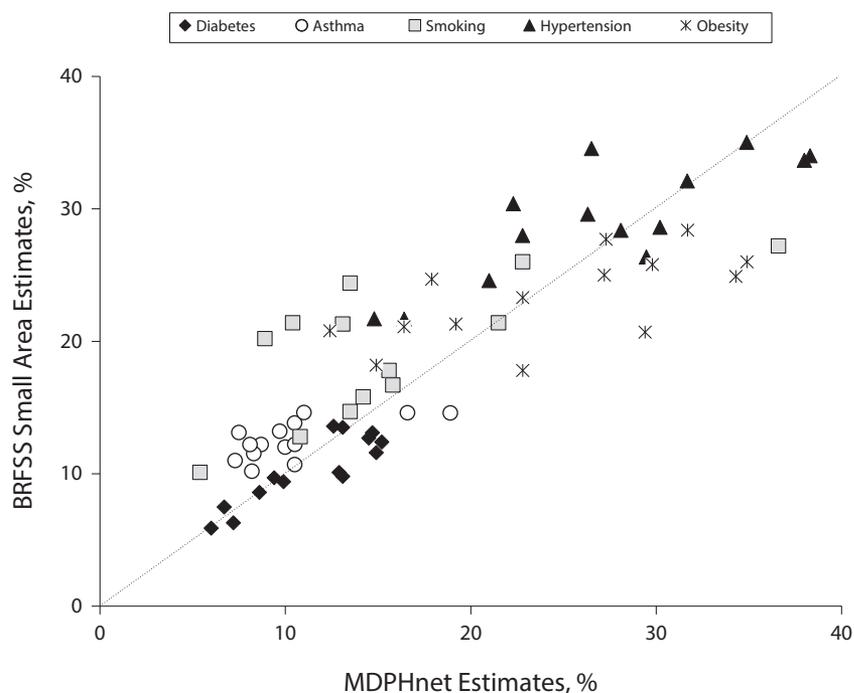
**TABLE 1—Diabetes, Asthma, Smoking, Hypertension, and Obesity Prevalence in Adults per MDPHnet (n = 1 073 545) vs the Behavioral Risk Factor Surveillance System (BRFSS) and 500 Cities Estimates (n = 15 654): Massachusetts, 2014**

Location	Census Population Aged ≥ 20 y, No.	MDPHnet Population Aged ≥ 20 y, No. (%)	Diabetes Prevalence, %			Asthma Prevalence, %			Smoking Prevalence, %			Hypertension Prevalence, %			Obesity Prevalence, %		
			MDPHnet Adjusted	BRFSS & 500 Cities	MDPHnet Adjusted	BRFSS & 500 Cities	MDPHnet Adjusted	BRFSS & 500 Cities	MDPHnet Adjusted	BRFSS & 500 Cities	MDPHnet Adjusted	BRFSS & 500 Cities	MDPHnet Adjusted	BRFSS & 500 Cities	MDPHnet Adjusted	BRFSS & 500 Cities	
Massachusetts	4 926 489	1 073 545 (21.8)	9.4	9.7	10.0	12.0	14.7	13.5	14.7	26.3	29.6	22.8	23.3				
Boston	482 004	138 369 (28.7)	8.6	8.6	8.7	12.2	17.8	15.6	17.8	21.0	24.6	19.2	21.3				
Brockton	67 018	24 983 (37.3)	14.5	12.7	10.5	13.8	26.0	22.8	26.0	34.9	35.0	31.7	28.4				
Cambridge	87 853	38 875 (44.3)	6.0	5.9	7.3	11.0	12.8	10.8	12.8	14.8	21.7	12.4	20.8				
Fall River	67 452	5 514 (8.2)	15.2	12.4	11.0	14.6	42.1	42.1	27.9	38.0	33.7	34.3	24.9				
Lawrence	51 375	1 458 (2.8)	12.6	13.6	7.5	13.1	10.4	10.4	21.4	22.3	30.4	27.3	27.7				
Lowell	76 952	26 420 (34.3)	13.1	9.8	10.5	12.2	8.9	20.2	20.2	29.5	26.4	29.4	20.7				
Lynn	65 129	34 110 (52.4)	14.9	11.6	9.7	13.2	21.5	21.5	21.4	31.7	32.1	17.9	24.7				
New Bedford	70 533	19 586 (27.8)	14.7	13.1	18.9	14.6	36.6	36.6	27.2	38.3	34.0	34.9	26.0				
Newton	62 002	15 505 (25.0)	6.7	7.5	8.2	10.2	5.4	5.4	10.1	22.8	28.0	14.9	18.2				
Quincy	75 266	32 695 (43.4)	9.9	9.4	10.5	10.7	15.8	15.8	16.7	28.1	28.4	22.8	17.8				
Somerville	64 498	40 792 (63.2)	7.2	6.3	8.3	11.5	14.2	14.2	15.8	16.4	21.6	16.4	21.1				
Springfield	105 158	410 (0.4)	13.1	13.5	16.6	14.6	13.5	13.5	24.4	26.5	34.5	29.8	25.8				
Worcester	132 434	22 161 (16.7)	12.9	10.1	8.1	12.2	13.1	13.1	21.3	30.2	28.6	27.2	25.0				
Pearson correlation coefficient for MDPHnet vs 500 Cities <sup>a</sup>																	
Crude			0.693			0.530			0.702			0.723					
Adjusted			0.890			0.672			0.747			0.828					

<sup>a</sup>Cities only.

estimating the denominator of patients under surveillance since there is no formal enrollment file for clinical practices as there is for insurance plans. There is also potential for MDPHnet to double-count patients that receive care in more than 1 MDPHnet-affiliated practice network. Furthermore, there may be local idiosyncrasies in practice patterns and data quality that introduce error into MDPHnet estimates (local estimates are more vulnerable to this problem than state-level estimates because local estimates often depend on data from a smaller number of practices). The problem of non-representative patient populations may be magnified in jurisdictions with lower rates of health insurance coverage than Massachusetts. Conversely, it may be possible to enhance the accuracy of EHR-based surveillance by being attentive to achieving minimum levels of population coverage, selecting clinical partners carefully to ensure geographic and demographic diversity, working with partners to increase data quality and consistency, and enhancing the disease projection framework to incorporate additional variables beyond age, gender, and race/ethnicity.

Possible sources of error in the 500 Cities small-area estimates could include the inherent limitations in trying to extrapolate disease rates using statistical models applied to small numbers of observations. It is notable in this regard that MDPHnet's sample sizes per city were in many cases substantially larger than the entire state-level sample for BRFSS. MDPHnet uses direct observation rather than imputation to estimate local disease rates. In addition, BRFSS disease estimates are based on self-reports, which may correlate poorly with clinical testing.<sup>25</sup> Patients may not be aware of their diagnoses, they may misunderstand their diagnoses, or they may misrepresent data on sensitive issues such as height, weight, and smoking. Finally, BRFSS assessments are based on questions that may not correlate with EHR-based measures of disease. For example, the BRFSS question, "Has a healthcare provider ever informed you that you have asthma?" could encompass a wide range of disease types (e.g., mild occasional wheezing, exercise-induced asthma, severe asthma) at various levels of disease activity (from fully resolved to currently flaring). The MDPHnet asthma case



Note. Each marker indicates a single location and condition. The diagonal dotted line marks where perfect agreement between MDPHnet and BRFSS would lie.

**FIGURE 2—Comparison of MDPHnet vs Behavioral Risk Factor Surveillance System (BRFSS) Small-Area Estimates: Massachusetts, 2014**

definition, by contrast, focuses on medically attended asthma. This may account for the lower observed asthma rates in MDPHnet than in the 500 Cities survey. It is notable that detailed evaluations of adults who self-report a diagnosis of asthma suggest that up to a third may not in fact have the disease.<sup>26,27</sup>

### EHR-Based Surveillance in Context

Determining whether EHR-based surveillance or BRFSS small-area estimates are more accurate would ideally require detailed in-person clinical evaluations of representative samples of patients in multiple communities to establish a reference standard against which to compare EHR-based surveillance and BRFSS small-area estimates. So long as EHR-based analyses are internally consistent, however, they should be able to facilitate rate comparisons between communities and across time. This may be adequate to help public health agents decide how to deploy resources and monitor the effects of interventions, particularly during urgent situations that require very timely data to inform

rapid responses, such as the current opioid crisis. This requires further evaluation.

Our data add to the growing literature on the potential value of EHR data for public health surveillance. New York City officials have shown that EHR-based estimates of city-level smoking and obesity rates compare favorably with those of New York City's Health and Nutrition Examination Survey and Community Health Survey.<sup>16</sup> Researchers in Boston have demonstrated that practice-based estimates of neighborhood smoking rates are similar to local BRFSS-based estimates.<sup>28</sup> And researchers in Denver, Colorado, have demonstrated that distributed analyses of EHR systems provide credible data on changes in obesity over time.<sup>29</sup>

Federal support for BRFSS has been declining. This has compelled states and cities either to supplement their BRFSS budgets with additional funds of their own or to decrease their sample size. Smaller sample sizes further constrain the utility of BRFSS for small-area and small-population surveillance. Health departments sometimes combine data across multiple years to increase sample size,

but this further delays the availability of results and blunts the capacity to evaluate temporal trends. Investing in EHR-based surveillance systems may ultimately prove to be a cost-effective complement to BRFSS and NHANES to monitor selected health outcomes. EHR-surveillance is unlikely, however, to ever replace BRFSS or NHANES. Some outcomes included in BRFSS and NHANES, particularly health risk behaviors and self-perceived health, are poorly and inconsistently documented in EHRs. In addition, clinical providers do not systematically test all patients for all conditions, and thus EHR-based surveillance has limited utility for estimating the frequency of under-diagnosed conditions such as diabetes and hypertension.<sup>30,31</sup>

### Public Health Implications

Our data suggest that EHR-based public health surveillance has the potential to generate state and local estimates of disease prevalence that approximate those of traditional sources. EHR-based surveillance can provide public health officials with timely, granular data about health in their communities that can be used to identify priority areas for intervention and to track the impact of public health initiatives. Further work is necessary to assess the generalizability of our findings to other jurisdictions, better define sources of accuracy and inaccuracy in EHR-based estimates, establish the minimal levels of population coverage and population diversity to generate accurate disease estimates, and develop financing and incentive models to facilitate the spread of EHR-based public health surveillance. 

### CONTRIBUTORS

M. Klompas, N. M. Cocoros, N. Nasuti, and T. Land conceptualized the project. M. Klompas, N. M. Cocoros, D. Erani, E. Hafer, B. Herrick, M. Josephson, M. Lee, M. D. Payne Weiss, B. Zambarano, K. R. Eberhardt, J. Malenfant, L. Nasuti, and T. Land designed and created the MDPHnet surveillance platform. M. Klompas, N. M. Cocoros, J. T. Menchaca, M. Josephson, B. Zambarano, and K. Eberhardt analyzed the data. L. Nasuti and T. Land obtained funding. M. Klompas drafted the article. All authors critically reviewed, revised, and approved the article.

### ACKNOWLEDGMENTS

Funding for this work was provided by the Massachusetts Department of Public Health.

B. Zambarano and K. R. Eberhardt are employees of and stockholders in Commonwealth Informatics, Inc., which is under contract to provide ongoing consulting

support for ESPHealth and PopMedNet software systems referenced in this article.

## HUMAN PARTICIPANT PROTECTION

MDPHnet was reviewed and approved by the Harvard Pilgrim Health Care Human Studies Committee with waiver of informed consent.

## REFERENCES

- Clark CR, Williams DR. Understanding county-level, cause-specific mortality: the great value—and limitations—of small area data. *JAMA*. 2016;316(22):2363–2365.
- Johnson NB, Hayes LD, Brown K, Hoo EC, Ethier KA, Centers for Disease Control and Prevention. CDC National Health Report: leading causes of morbidity and mortality and associated behavioral risk and protective factors—United States, 2005–2013. *MMWR Suppl*. 2014; 63(4):3–27.
- Chowdhury PP, Mawokomatanda T, Xu F, et al. Surveillance for certain health behaviors, chronic diseases, and conditions, access to health care, and use of preventive health services among states and selected local areas—Behavioral Risk Factor Surveillance System, United States, 2012. *MMWR Surveill Summ*. 2016;65(4):1–142.
- Guo F, Garvey WT. Trends in cardiovascular health metrics in obese adults: National Health and Nutrition Examination Survey (NHANES), 1988–2014. *J Am Heart Assoc*. 2016;5(7):e003619.
- Mokdad AH. The Behavioral Risk Factors Surveillance System: past, present, and future. *Annu Rev Public Health*. 2009;30:43–54.
- Institute of Medicine. Existing surveillance data sources and systems. In: *A Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung Diseases*. Washington, DC: National Academies Press; 2011:65–90.
- Pierannunzi C, Hu SS, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Med Res Methodol*. 2013;13:49.
- Menke A, Casagrande S, Geiss L, Cowie CC. Prevalence of and trends in diabetes among adults in the United States, 1988–2012. *JAMA*. 2015;314(10):1021–1029.
- Flegal KM, Kruszon-Moran D, Carroll MD, Fryar CD, Ogden CL. Trends in obesity among adults in the United States, 2005 to 2014. *JAMA*. 2016;315(21):2284–2291.
- Klompas M, McVetta J, Lazarus R, et al. Integrating clinical practice and public health surveillance using electronic medical record systems. *Am J Public Health*. 2012;102(suppl 3):S325–S332.
- Vogel J, Brown JS, Land T, Platt R, Klompas M. MDPHnet: secure, distributed sharing of electronic health record data for public health surveillance, evaluation, and planning. *Am J Public Health*. 2014;104(12):2265–2270.
- Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health Aff (Millwood)*. 2014; 33(7):1178–1186.
- Eggleston EM, Weitzman ER. Innovative uses of electronic health records and social media for public health surveillance. *Curr Diab Rep*. 2014;14(3):468.
- Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health*. 2015;36:345–359.
- Newton-Dame R, McVeigh KH, Schreibstein 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.
- McVeigh KH, Newton-Dame R, Chan PY, et al. Can electronic health records be used for population health surveillance? Validating population health metrics against established survey data. *EGEMS (Wash DC)*. 2016;4(1):1267.
- Romo ML, Chan PY, Lurie-Moroni E, et al. Characterizing adults receiving primary medical care in New York City: implications for using electronic health records for chronic disease surveillance. *Prev Chronic Dis*. 2016;13:E56.
- Lazarus R, Klompas M, Campion FX, et al. Electronic support for public health: validated case finding and reporting for notifiable diseases using electronic medical data. *J Am Med Inform Assoc*. 2009;16(1):18–24.
- Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care*. 2010;48(6 suppl):S45–S51.
- Klann JG, Buck MD, Brown J, et al. Query Health: standards-based, cross-platform population health surveillance. *J Am Med Inform Assoc*. 2014; 21(4):650–656.
- Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care*. 2013;36(4):914–921.
- Centers for Disease Control and Prevention. 500 Cities: local data for better health. 2017. Available at: <https://www.cdc.gov/500cities/index.htm>. Accessed January 16, 2017.
- Zhang X, Holt JB, Lu H, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *Am J Epidemiol*. 2014;179(8):1025–1033.
- Zhang X, Holt JB, Yun S, Lu H, Greenlund KJ, Croft JB. Validation of multilevel regression and post-stratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *Am J Epidemiol*. 2015;182(2):127–137.
- Connor Gorber S, Tremblay M, Moher D, Gorber B. A comparison of direct vs. self-report measures for assessing height, weight and body mass index: a systematic review. *Obes Rev*. 2007;8(4):307–326.
- Holm M, Omenaas E, Gislason T, et al. Remission of asthma: a prospective longitudinal study from northern Europe (RHINE study). *Eur Respir J*. 2007;30(1):62–65.
- Aaron SD, Vandemheen KL, FitzGerald JM, et al. Reevaluation of diagnosis in adults with physician-diagnosed asthma. *JAMA*. 2017;317(3):269–279.
- Linder JA, Rigotti NA, Brawarsky P, et al. Use of practice-based research network data to measure neighborhood smoking prevalence. *Prev Chronic Dis*. 2013;10: E84.
- Davidson AJ, McCormick EV, Dickinson LM, Haemer MA, Knierim SD, Hambidge SJ. Population-level obesity surveillance: monitoring childhood body mass index z-score in a safety-net system. *Acad Pediatr*. 2014;14(6):632–638.
- Leong A, Dasgupta K, Chiasson JL, Rahme E. Estimating the population prevalence of diagnosed and undiagnosed diabetes. *Diabetes Care*. 2013;36(10): 3002–3008.
- Holt TA, Gunnarsson CL, Clod PA, Ross SD. Identification of undiagnosed diabetes and quality of diabetes care in the United States: cross-sectional study of 11.5 million primary care electronic records. *CMAJ Open*. 2014;2(4):E248–E255.