

RESEARCH

Open Access



Equivalence of electronic health record data for measuring hypertension prevalence: a retrospective comparison to BRFSS with data from two Indiana health systems, 2021

Katie S. Allen^{1,2*}, Justin Stiles^{1,2}, Veronica M. Daye³, Ashley Wiensch¹, Nimish Valvi⁴ and Brian E. Dixon^{1,5}

Abstract

Background Public health surveillance requires timely access to actionable data at every level. Current approaches for accessing chronic disease surveillance data are not sufficient, and health departments are increasingly looking to augment surveillance efforts using electronic health records (EHRs). While proven effective for acute syndromic surveillance, the utilization of EHR systems and health data networks for monitoring chronic conditions remains sparse. This study tested the generalizability of a previously validated hypertension computable phenotype.

Methods A previously developed phenotype was used to estimate prevalence of hypertension in a geographically and clinically distinct region from its development. To test validity, the results were compared to available, statewide Behavioral Risk Factor Surveillance System (BRFSS) data using the two one-sided t-test (TOST) of equivalence between BRFSS- and EHR-based prevalence estimates. The TOST was performed at the overall level as well as stratified by age, gender, and race/ethnicity.

Results Compared to statewide hypertension prevalence of 34.5% in the BRFSS, an EHR-based phenotype estimated an overall prevalence of 24.1%. Estimates were not equivalent overall or across most subpopulations. Like BRFSS, we observed higher prevalence among Black men and women as well as increasing prevalence with age.

Conclusion With caveats, this study demonstrates that EHR-derived prevalence estimates may serve as a complement for population-based survey estimates. Utilizing available EHR data should increase timeliness of surveillance as well as enhance the ability of states and local health agencies to more readily address the burden of chronic disease in their respective jurisdictions.

Keywords Public health surveillance, Public health, Chronic conditions

*Correspondence:

Katie S. Allen
allenkat@regenstein.org

¹Regenstrief Institute, Inc, 1101 W. 10th Street, Indianapolis, IN 46202, USA

²Indiana University School of Medicine, 1101 W. 10th Street, Indianapolis, IN 46202, USA

³Indiana Department of Health, 2 Meridian Street, Indianapolis, IN 46204, USA

⁴Ball State University, Health Professionals Building (HB 350), Muncie, IN 47306, USA

⁵Fairbanks School of Public Health, Indiana University, 1101 W. 10th Street, Indianapolis, IN 46202, USA



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

While efforts to enhance the public health infrastructure have been underway for almost a decade, inadequacies uncovered by the COVID-19 pandemic prompted targeted action. The Centers for Disease Control and Prevention (CDC) is facilitating a broad set of data modernization (DMI) activities at federal, state, and local levels [1]. A major focus in these DMI efforts is leveraging clinical data already captured in electronic health record (EHR) systems and health information exchange (HIE) networks [2]. Use of EHR systems for public health surveillance of infectious diseases, including SARS-CoV-2 and sexually transmitted diseases, has become routine in many jurisdictions [3, 4].

While proven useful for infectious disease surveillance, the utilization of EHR and HIE technologies for monitoring chronic conditions remains sparse. Enhancing chronic disease surveillance systems and processes can improve disease surveillance by identifying populations at risk. Additionally, it can assist with effective implementation of public health interventions for primary and secondary prevention. However, identification of chronic conditions within EHR data requires a validated, disease-specific phenotype to ensure accuracy [5]. Phenotypes for EHR-derived data have previously been developed and implemented for conditions including diabetes mellitus [6] and opioid use disorder [7]. Additional phenotypes are needed for other critical chronic conditions.

Hypertension is a chronic condition that is a risk factor for heart disease and stroke, which are two of the leading causes of death in the United States [8, 9]. Accordingly, preventing and controlling hypertension would result in large public health benefits. Given the importance of hypertension surveillance, creating and validating an EHR-derived phenotype for hypertension is a priority. There are several initiatives underway to address this issue [10, 11] but testing the generalizability of these phenotypes remains an understudied area, partially given the difficulty related to sharing of clinical data.

One project has addressed the generalizability of phenotypes by creating a network focused on this important work. The Multi-State EHR-Based Network for Disease Surveillance (MENDS) pilot project is a national network focused on chronic disease surveillance [12, 13]. As part of these efforts, MENDS developed an EHR-derived phenotype for identification of patients with hypertension [14]. This phenotype leverages blood pressure values, hypertensive medications, and diagnosis codes to generate a prevalence estimate. The phenotype has previously been compared to clinical data derived from a single institution in Louisiana, but a larger generalizability study has not yet been conducted [14].

The purpose of this study was to test the generalizability of the MENDS hypertension phenotype by

implementing it at another geographically and clinically distinct [12] region. This region represents two multi-hospital health systems and a larger patient population. The overall goal of this study is to test generalizability of EHR-derived hypertension phenotypes as a step towards adoption of more-timely chronic disease surveillance.

Methods

Data sources

Indiana University and Regenstrief Institute joined the MENDS network in 2020, which enabled sharing of clinical data and enhanced research capabilities. The data shared with MENDS is extracted via a regional health information exchange (HIE). Through participation in MENDS [12], the Regenstrief Institute shares data from two of the larger health systems in Indiana. Together, these health systems represent 1.5 million patients, approximately 22% of the population of the state, and multiple counties within the State. Between the two systems, millions of outpatient visits are recorded annually.

Available data are derived from EHR systems connected to the Indiana Network for Patient Care (INPC), a statewide HIE network in Indiana. [15] Shared EHR data cover emergency department visits, hospital admissions, and large outpatient healthcare clinics. Data were extracted from the two health systems for adults (at least 18 years of age) with a home address in the state. For this analysis, the population represents those individuals who sought care at either of the two health systems between January 1, 2020 and December 31, 2021.

The MENDS phenotype [14] utilizes two years of EHR data to capture a representative number of clinical encounters, since individual healthcare utilization may not occur annually. This is especially true for individuals who consider themselves to be healthy. The MENDS hypertension algorithm was run against the Indiana EHR-derived dataset. This data is made available in the form of population-level statistics via the MENDS RiskScape data visualization portal [16] which is provided to all MENDS sites. This portal serves as a method to display and interact with calculated population-level estimates. Accordingly, the analysis for this study is done utilizing aggregated data retrieved from RiskScape on October 2, 2023 and covering the time period specified above.

We sought to compare publically available demographics to the EHR-derived demographics. To establish comparisons across other data categories, demographic data was procured from the Behavioral Risk Factor Surveillance System (BRFSS), which is a nationally implemented survey of health-related behaviors, chronic health conditions, and use of preventive services in the U.S. For this study, demographic variables from BRFSS included: population size, age groups, and race/ethnicity data.

For comparison of hypertension results, we leveraged BRFSS. The prevalence estimates produced by BRFSS are carefully developed, validated, and are weighted to minimize biases in response or coverage.⁽¹⁷⁾ BRFSS collects data in all 50 states, the District of Columbia, and territories. However, for small geographics (e.g., county) or population subgroups, the large confidence intervals of BRFSS estimates may suggest imprecision.⁽¹⁸⁾ For this study, we utilized statewide, Indiana data related to 2021 prevalence estimates for hypertension.

Measures

Demographic prevalence measures included the categories of: gender, race/ethnicity, and age groupings (18–24, 25–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80+). For EHR-derived data, age at the time of index was utilized for age group classification. The age groupings were selected as they were the shared groupings across BRFSS and MENDS data, facilitating comparison. Race and ethnicity are presented as a single combined variable, reflecting limitations in comparable data capture across MENDS and BRFSS. MENDS prevalence measures were retrieved from the RiskScape portal, which generates pre-calculated, population-level measurements. The MENDS hypertension algorithm utilizes a combination of diagnosis codes, antihypertensive medications, and blood pressure measurements. Blood pressure was considered elevated if ≥ 2 blood pressure values of $\geq 140/90$ were present on separate days [14].

Weighted BRFSS prevalence measures were compared to EHR-based weighted measures extracted from the Indiana MENDS dataset. The 2021 BRFSS facilitated an overall hypertension prevalence as well as rates by age, race/ethnicity, and gender. Hypertension prevalence was calculated utilizing the variable which asks the respondent “Have you even been told by a doctor, nurse or other health professional that you have high blood pressure?” Respondents indicating “Yes” but not noting gestationally related were considered to have hypertension. All other responses were considered negative. Given the uncertainly related to hypertension status, those who declined to answer the question (unknown or refused) were dropped from the analysis ($n = 972$, 9.9% of BRFSS cohort). Weighting was applied leveraging BRFSS’s weighting data and suggested methodologies, which accounts for population demographics and non-response [19].

Analysis

Descriptive statistics of demographics (age, gender, race/ethnicity) were created for both the BRFSS and clinical cohorts. They were compared utilizing absolute differences. The primary analysis consisted of an equivalency test between the MENDS-derived prevalence

and the BRFSS-derived prevalence. Equivalence testing examines whether two independent statistics are similar enough to be treated as though they are equivalent. The null hypothesis is that the statistics differ by at least a specified amount. If the test results in a p -value < 0.05 , then the null hypothesis is rejected with a conclusion that the two statistics differ by less than the specified amount. We employed the two one-sided t-test (TOST) to test equivalence between BRFSS- and MENDS-based prevalence estimates. The TOST has been utilized successfully in similar studies [20, 21] and is increasingly a tool for understanding equivalency in prevalence for public health data. Additionally, TOST methodology assists to address concerns related to large sample sizes, such as those obtained from EHR data [22]. The TOST was performed at the overall level as well as stratified by age, gender, race/ethnicity, age-gender, and gender-race/ethnicity. Analyses were conducted at the 95% confidence interval and at the 90% confidence interval, which has been determined to be a suitable interval for comparing population based estimates [20].

All analyses were conducted in R (4.2.1), with package “dplyr” and Microsoft Excel [15]. Institutional Review Board (IRB) approval for the MENDS network was approved by Indiana University. IRB approval for this specific study was not required analysis given the utilization of publicly available (BRFSS) and de-identified data.

Results

Descriptive statistics of demographics for the hypertensive group within each cohort (BRFSS and Clinical) are presented in Table 1. Overall, there are small differences between the two populations. The clinical data had more females compared to the BRFSS data (54.4% vs. 51.4% respectively). The age group distribution varied. Clinical data had slightly higher proportions of younger age, whereas BRFSS had slightly higher proportions for older populations. Black patients represented a greater proportion of patients in the clinical data compared to BRFSS (13.9% compared to 10.18%). The BRFSS cohort had more unknown/other race/ethnicity (3.78% compared to 1.38%).

The primary analysis, consisting of equivalency testing for prevalence for BRFSS compared those obtained from the MENDS RiskScape visualization portal, are presented in Table 2. The overall hypertension rates were 34.5% (BRFSS) and 24.1% (EHR) respectively. Both cohorts identified a higher prevalence of hypertension for males (BRFSS 35%, MENDS 24.5%) and increasing prevalence with age. TOST results showed wide absolute differences observed across many categories. No results were statistically significant at the 95% confidence interval level. At the 90% confidence interval, Hispanic ethnicity,

Table 1 Demographics and comparison of the clinically derived cohort and the hypertensive BRFSS cohort, state of Indiana, 2021

	BRFSS Cohort (n = 4,124)		Clinical Cohort (n = 301,124)		Difference
Gender					
Male	2,003	(48.57%)	137,435	(45.64%)	2.93%
Female	2,121	(51.43%)	163,689	(54.36%)	-2.93%
Age					
18–24	32	(0.78%)	3,525	(1.17%)	-0.39%
25–29	55	(1.33%)	7,378	(2.45%)	-1.12%
30–39	209	(5.07%)	24,296	(8.07%)	-3.00%
40–49	342	(8.29%)	37,569	(12.48%)	-4.18%
50–59	719	(17.43%)	59,972	(19.92%)	2.39%
60–69	1,135	(27.52%)	75,669	(25.13%)	4.94%
70–79	1,032	(25.02%)	60,485	(20.09%)	1.35%
80+	497	(12.05%)	32,230	(10.70%)	2.66%
Race					
White	3,417	(82.86%)	241,495	(80.20%)	2.66%
Black	29	(0.70%)	41,925	(13.92%)	-3.74%
Asian	420	(10.18%)	3,363	(1.12%)	-0.41%
Hispanic	102	(2.47%)	9,987	(3.32%)	-0.84%
Other	156	(3.78%)	4,153	(1.38%)	2.40%

aged 18–24, and female aged 18–24 were equivalent. The remaining were not statistically significant.

Discussion

Chronic diseases are major sources of morbidity and mortality and effective public health intervention to address these health burdens is critical [23]. In order to properly conduct surveillance and intervention activities, population health management requires timely access to data. Under the CDCs data modernization activities [1], clinically-derived data sources are a priority. Clinical data sets being collected through routine, real-time health care, could strengthen the ability of public health to monitor chronic conditions [24, 25]. Previous work by MENDS has advanced the ability to identify patients with chronic conditions, including hypertension, from within clinical data [14]. This study sought to advance these efforts by testing the generalizability of a previously proven phenotype.

Currently, public health surveillance for chronic disease is largely handled through the use of self-report data, such as the BRFSS. However, these large, manually intensive surveys require considerable resources and are often delayed due to the time required to field the survey and analyze results. Moreover, federal datasets are vulnerable to political influence, exemplified by BRFSS datasets taken offline on January 31, 2025 [26], and not yet online as of submission of this manuscript. Our study demonstrates that EHR-data may be a suitable complement to the currently utilized self-report data, especially if these data are captured and retained by local and state jurisdictions. For demographics, there are no meaningful differences, except for two categories: age group 40–49 and the Black

racial group, which were 4.18% and 3.74% higher in the EHR-derived data. For the Indiana Black community, the EHR-derived hypertension prevalence rate better aligns with known prevalence of the condition within the State. This suggests that EHR-derived estimates may be helpful for addressing minority populations within a community, which are often not well-represented in national surveys [27].

EHR-derived estimates may be complementary [28] to clinical data in several ways. First, clinical data estimates may be timelier than the national surveys (i.e., BRFSS). Locally representative, weighted surveys require time to administer and analyze the data, which result in delays that which may not reflect the current community trends. Adding to the delay concerns is that certain questions are frequently not asked each year, as is the case with BRFSS and hypertension. Instead, the hypertension question is asked every other year to maximize the number of surveillance items without overburdening the participant. The result is that these conditions cannot be tracked on a year-to-year basis. Additionally, the questions lack the specificity that is possible in clinically-derived data. As an example, BRFSS asks whether an individual “has ever been diagnosed” which reflects a lifetime, not necessarily current. EHR-based data can also include medications, which can allow for treatment prevalence or ascertaining controlled versus uncontrolled hypertension.

In addition to being more timely than national surveys, utilizing hypertension measurements which rely on EHR data may be more accurate in two ways. Calculations generated on localized, clinical data produce smaller confidence intervals, which may ensure stability and precision in the estimates over time. Moreover, national

Table 2 Results of TOST equivalency testing between the clinically derived cohort and the hypertensive BRFSS cohort, state of Indiana, 2020–2021

	BRFSS	Clinical			
	%	%	Δ	$\Delta 95\% \text{ CI}$	$\Delta 90\% \text{ CI}$
Overall	34.50%	24.10%	-10.4	(-11.5, 1.7)	(-11.3, 1.7)
Male	37.20%	24.50%	-12.7	(-14.4, 2.5)	(-14.1, 2.3)
Female	32.00%	23.80%	-8.2	(-9.7, 2.2)	(-9.4, 2.0)
Caucasian	35.80%	26.30%	-9.5	(-10.7, 1.9)	(-10.5, 1.7)
Asian	14.00%	15.30%	1.3	(-4.8, 9.2)	(-3.8, 8.2)
Black	42.80%	27.40%	-15.4	(-19.8, 6.6)	(-19.0, 5.9)
Hispanic	15.60%	15.30%	-0.3	(-3.8, 5.3)	(-3.2, 4.7)*
Other	28.00%	4.70%	-23.3	(-28.8, 8.3)	(-27.9, 7.4)
18–24	6.05%	4.70%	-1.4	(-3.7, 3.5)	(-3.3, 3.1)*
25–29	13.00%	6.50%	-6.5	(-10.3, 5.7)	(-9.7, 5.1)
30–39	18.00%	11.10%	-6.9	(-9.4, 3.8)	(-9.0, 3.4)
40–49	27.10%	18.90%	-8.2	(-11.0, 4.2)	(-10.5, 3.8)
50–59	42.20%	28.70%	-13.5	(-16.2, 4.0)	(-15.7, 3.6)
60–69	56.70%	36.70%	-20.0	(-22.5, 3.8)	(-22.1, 3.4)
70–79	66.00%	42.80%	-23.2	(-26.0, 4.2)	(-25.5, 3.7)
80+	62.20%	37.80%	-24.4	(-28.7, 6.5)	(-28.0, 5.8)
Male Caucasian	38.61%	27.00%	-11.6	(-13.5, 2.8)	(-13.2, 2.5)
Male Asian	16.04%	17.20%	1.2	(-7.7, 13.3)	(-6.2, 11.9)
Male Black	45.59%	24.80%	-20.8	(-27.7, 10.4)	(-26.6, 9.3)
Male Hispanic	17.17%	15.80%	-1.4	(-6.5, 7.8)	(-5.7, 6.9)
Male Other	31.09%	5.40%	-25.7	(-33.4, 11.7)	(-32.2, 10.4)
Female Caucasian	33.11%	25.70%	-7.4	(-9.0, 2.5)	(-8.8, 2.2)
Female Asian	11.95%	14.00%	2.1	(-6.2, 12.5)	(-4.9, 11.2)
Female Black	40.31%	29.40%	-10.9	(-16.3, 8.2)	(-15.4, 7.3)
Female Hispanic	13.91%	14.80%	0.9	(-3.9, 7.2)	(-3.1, 6.4)
Female Other	24.55%	4.10%	-20.4	(-27.8, 11.2)	(-26.6, 10.0)
Male 18–24	8.60%	4.20%	-4.4	(-8.2, 5.8)	(-7.6, 5.1)
Male 25–29	14.78%	5.70%	-9.1	(-14.8, 8.7)	(-13.9, 7.7)
Male 30–39	22.09%	10.60%	-11.5	(-15.4, 5.9)	(-14.8, 5.3)
Male 40–49	32.91%	19.30%	-13.6	(-18.0, 6.6)	(-17.3, 5.1)
Male 50–59	45.80%	29.60%	-16.2	(-20.1, 5.9)	(-19.5, 5.2)
Male 60–69	60.42%	37.20%	-23.2	(-26.8, 5.4)	(-26.2, 4.8)
Male 70–79	70.29%	42.30%	-28.0	(-31.9, 5.8)	(-31.2, 5.2)
Male 80+	60.05%	37.60%	-22.4	(-29.6, 10.8)	(-28.4, 9.6)
Female 18–24	3.36%	5.00%	1.6	(-0.9, 3.8)	(-0.5, 3.4)*
Female 25–29	11.18%	7.00%	-4.2	(-9.1, 7.4)	(-8.3, 6.6)
Female 30–39	13.93%	11.40%	-2.5	(-5.6, 4.6)	(-5.1, 4.1)
Female 40–49	21.64%	18.60%	-3.0	(-6.5, 5.3)	(-6.0, 4.7)
Female 50–59	38.58%	27.90%	-10.7	(-14.3, 5.4)	(-13.7, 4.8)
Female 60–69	53.18%	36.20%	-17.0	(-20.5, 5.3)	(-19.9, 4.7)
Female 70–79	62.40%	43.20%	-19.2	(-23.1, 5.8)	(-22.4, 5.2)
Female 80+	63.53%	38.00%	-25.5	(-31.0, 8.3)	(-30.1, 7.4)

*Results statistically significant

surveys, like BRFSS, target representation at national and state levels and are, therefore, not always locally representative [29]. Additionally, the BRFSS also suffers from poor response rates [27], which may influence selection bias. This is particularly important for jurisdictions with subpopulations that may not be represented in national surveys (e.g., American Indians, refugees from a certain

region). This is of critical importance, especially given the imperative for improving health outcomes across all populations. Clinical data represent patients being seen locally for care, enabling the highly localized estimates and therefore more actionable by local jurisdictions. Previous work [30] demonstrates that EHR-based data can be used to estimate prevalence at the neighborhood

level, enabling better stewardship of local public health resources.

Another way that clinical data may be preferable compared to self-reported survey information relates to response and recall bias. The clinical phenotypes leverage variables (blood pressure measurements, diagnosis codes) that are collected at the point of care, rather than relying on the recall of survey respondents. Patient recall can be influenced by many phenomena, such as recency of the information being obtained, the respondents mood at the time of questioning, and their past experiences related to the condition [31]. Comparatively, clinical data are captured at a point in time by a health care provider and represent rates of hypertension during the period of interest. For example, BRFSS asks whether the individual has been told they are hypertensive in their lifetime, which may not reflect their current health status. However, the phenotypes for estimating clinical prevalence can cause variation and public health jurisdictions should seek to understand the data thoroughly and think about implications for interpretation. Results can vary depending on the phenotype utilized to identify hypertensive patients [11], the availability of data elements (e.g., blood pressure measurements), or other elements such as geography or age-standardization [32]. In this study, the BRFSS data represented a statewide population versus the EHR data from two large providers representing less than a quarter of the state population. This difference likely accounts for the non-significance of findings, yet the estimates might still represent truth given they were measured from real patients presenting for care.

An additional benefit of leveraging clinical data is the ability to adjust to current clinical guidelines. Self-reported survey data lacks the specificity to understand whether the diagnosis falls within current clinical guidelines. For example, self-report data is not sensitive to the recent shift to classifying an individual as hypertensive at $\geq 130/80$ compared to the previous standard of $\geq 140/90$. Self-reported data may lack nuance where clinical phenotypes can be updated to address these evolving definitions.

There are considerations to be aware of when proceeding with clinically derived data. As with any calculations of prevalence, there are particular features to pay close attention to. A prime example for represented in this study is that the timeframe overlaps the time period of COVID-19, where healthcare patterns varied, with delayed care being a primary concern [33]. Accordingly, the prevalence produced may not reflect the true rate of hypertension in the community for the time period. Additionally, the clinical population may differ from survey populations. For example, the clinical population are more likely to have chronic conditions, be experiencing acute symptoms (opposed to asymptomatic), be older,

and insured [34–36]. Moreover, in this study, the clinical population represented around one-quarter of the population and geography of the state compared with BRFSS which sampled individuals across the state (although we note many counties are under-represented in the BRFSS). When calculating prevalence it is important to understand all parameters, including data collection parameters, that may have existed during the period of interest. Weighting and adjusting for sampling is critical to incidence and prevalence estimates using any data source. Moreover, examining longitudinal data and trends across years should also be done in the future to better assess how EHR data can complement other data sources.

The widespread use of EHR-based data for surveillance is not without limitations. Importantly, barriers currently exist related to data access. Many data desired by public health are not currently part of mandatory reporting requirements at local or national levels, and governance barriers often exist for direct access by local or state health departments. To-date, mandatory reporting laws have focused on communicable diseases, making surveillance of chronic conditions utilizing clinically-derived data challenging and buy-in difficult [2]. It is posited that addressing these barriers is critical to a modernized data infrastructure and it should be a priority to facilitate data-sharing activities at state and local levels [2]. There are several examples of successful data-sharing governance frameworks, including the one used by the MENDS network [12]. In addition, the U.S. Office of the National Coordinator for Health Information Technology has certified several Qualified Health Information Networks or QHINs that leverage a common framework [37] for trusted exchange of clinical data, which the U.S. government hopes can also be leveraged for public health use cases [38]. The public health community should leverage existing models and lessons learned, where possible, to enable data sharing for chronic disease surveillance.

While EHR-derived data has many benefits for surveillance, it is important that the public health community understand its limitations. A major concern is data quality and accuracy [39]. Large, clinically-derived data sets can produce statistically significant results given the sheer number of observations [40]. However, data quality issues may influence the accuracy of these results. Understanding the data being utilized to generate the phenotype is critical. Additionally, careful consideration of the derived prevalence should be undertaken. While results may not be statistically different, the absolute numbers may be meaningful.

This observational study has limitations worth noting. First, observational studies leveraging real-world clinical data are subject to bias and confounding factors that are not present in other research methods [41]. One such factor is that the population seeking clinical care may be

different from the general population in ways that are not fully understood. This study examined the demographics and found only minor differences in the representativeness of the two cohorts, but this may not account for variation based upon health status, akin to the limitations of healthcare seeking behavior discussed above. The limited diversity of Indiana may limit the generalizability to other areas and may have inhibited the detection of meaningful race/ethnicity differences. This study also has limitations related to geography and data collection. Patients come from multiple counties around the entire State, but this may not be representative of the state as a whole. This study is unable to analyze the prevalence changes with the updated clinical guidelines, which may affect the long-term equivalence of the phenotype. However, both data sets were gathered prior to this change being implemented. Additionally, this study leverages a comparison of a survey question (“have you ever in your lifetime...”) compared to clinical measurements, which may influence the findings of this study. Another limitation is the analyses involving racial and ethnic data suffer from missing or poor data collection [42], which is a key concern critical to address as part of data modernization efforts.

Conclusion

Utilizing clinical data for chronic disease surveillance is a high-value opportunity for public health systems, yet it is currently infrequently leveraged or not leveraged at its full capability. The CDC’s data modernization efforts include leveraging data sources, such as clinical EHR data related to chronic diseases, to enable action at combating morbidity and mortality. The MENDS project uniquely contributes to these activities by creating robust governance and methods for supporting surveillance of chronic disease. Specifically, the development of validated, reusable phenotypes is a critical component towards scaling of chronic disease surveillance using EHR-derived data. This project has shown that phenotypes developed and tested in other geographic locations are generalizable beyond their population borders and may be a complementary tool for chronic disease surveillance. Sharing these phenotypes and lessons will support national surveillance efforts to better measure and address chronic disease.

Abbreviations

BRFSS	Behavioral Risk Factor Surveillance System
CDC	Centers for Disease Control and Prevention
DMI	Data Modernization Initiative
EHR	Electronic Health Record
HIE	Health Information Exchange
INPC	Indiana Network for Patient Care
IRB	Institutional Review Board
MENDS	Multi-State EHR-Based Network for Disease Surveillance
TOST	Two-One Sided T-test

Acknowledgements

We acknowledge the contribution of MENDS partner sites and project team that participated in the creation of this information (<https://chronicdisease.org/page/MENDSinfo/>). The authors further thank Commonwealth Informatics and the Regenstrief Data Services teams for their work to extract and transform EHR data for the MENDS Project. Additionally, our thanks go to Laura Heinrich and Michael Ramey, Jr. at the Indiana Department of Health and Andrea Bochenek at the Marion County Health Department for their support.

Author contributions

B.D. and J.S. conceived of this project; N.V., J.S., and K.A. completed analytical work; V.D. provided subject matter expertise and review; A.W. provided administrative support; K.A. drafted the manuscript; all authors provided manuscript review and feedback.

Funding

The “Improving Chronic Disease Surveillance and Management Through the Use of Electronic Health Records/Health Information Systems” project is supported by the Centers for Disease Control and Prevention (CDC) of the US Department of Health and Human Services (HHS) as part of a financial assistance award totaling \$1,800,000 with 100% funded by CDC/HHS. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, CDC/HHS, or the US Government.

Data availability

All data generated during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study was reviewed by the Indiana University Institutional Review Board (IRB) and deemed non-human subjects. Consent to participate was also waived by the IRB. Only aggregate data was utilized for the analyses presented within this manuscript and therefore no individual persons data is included. This research adhered to the Declaration of Helsinki. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication

N/A.

Competing interests

The authors declare no competing interests.

Received: 2 January 2024 / Accepted: 20 March 2025

Published online: 04 April 2025

References

1. The Centers for Disease Control and Prevention (CDC). Data Modernization Initiative [Internet]. Better Data. Better Decisions. Better Health. 2022. Available from: <https://www.cdc.gov/surveillance/data-modernization/index.html>
2. Acharya JC, Staes C, Allen KS, Hartsell J, Cullen TA, Lenert L et al. Strengths, weaknesses, opportunities, and threats for the nation’s public health information systems infrastructure: synthesis of discussions from the 2022 ACMI Symposium. *J Am Med Inform Assoc*. 2023;30(6):1011–21.
3. Dixon BE, Grannis SJ, McAndrews C, Broyles AA, Mikels-Carrasco W, Wiensch A, et al. Leveraging data visualization and a statewide health information exchange to support COVID-19 surveillance and response: application of public health informatics. *J Am Med Inform Assoc*. 2021;28(7):1363–73.
4. Dixon BE, Zhang Z, Arno JN, Revere D, Joseph Gibson P, Grannis SJ. Improving notifiable disease case reporting through electronic information Exchange-Facilitated decision support: A controlled Before-and-After trial. *Public Health Rep*. 2020;135(3):401–10.
5. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inf Assoc*. 2014;21(2):221–30.
6. Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P, Carroll RJ, et al. A case study evaluating the portability of an executable computable

- phenotype algorithm across multiple institutions and electronic health record environments. *J Am Med Inform Assoc.* 2018;25(11):1540–6.
7. Chartash D, Paek H, Dziura JD, Ross BK, Noguee DP, Boccio E, et al. Identifying opioid use disorder in the emergency department: Multi-System electronic health Record-Based computable phenotype derivation and validation study. *JMIR Med Inf.* 2019;7(4):e15794.
8. World Health Organization [Internet]. [cited 2023 Jan 15]. Key Facts About Hypertension. Available from: <https://www.who.int/news-room/fact-sheets/detail/hypertension>
9. Merai R, Siegel C, Rakotz M, Basch P, Wright J, Wong B, et al. CDC grand rounds: A public health approach to detect and control hypertension. *MMWR Morb Mortal Wkly Rep.* 2016;65(45):1261–4.
10. Horth RZ, Wagstaff S, Jeppson T, Patel V, McClellan J, Bissonette N, et al. Use of electronic health records from a statewide health information exchange to support public health surveillance of diabetes and hypertension. *BMC Public Health.* 2019;19(1):1106.
11. Valvi NR, Allen KS, Gibson PJ, McFarlane TD, Dixon BE. Local Health Department Surveillance of Hypertension using Electronic Health Record Data. Under Review.
12. Kraus EM, Saintus L, Martinez AK, Brand B, Begley E, Merritt RK et al. Fostering Governance and Information Partnerships for Chronic Disease Surveillance: The Multi-State EHR-Based Network for Disease Surveillance. *Journal of Public Health Management and Practice* [Internet]. 2023 Oct 5 [cited 2023 Dec 4]; Available from: <https://journals.lww.com/https://doi.org/10.1097/PHH.0000000000001810>
13. Hohman KH, Martinez AK, Klompas M, Kraus EM, Li W, Carton TW, et al. The Multi-State EHR-Based Network for Disease Surveillance. *J Public Health Manage Pract.* 2023;29(2):162–73. Leveraging Electronic Health Record Data for Timely Chronic Disease Surveillance.
14. Hohman KH, Zambarano B, Klompas M, Wall HK, Kraus EM, Carton TW, et al. Development of a hypertension electronic phenotype for chronic disease surveillance in electronic health records: key analytic decisions and their effects. *Prev Chronic Dis.* 2023;20:230026.
15. Overhage MJ, Kinsky J. The Indiana health information exchange. In: Dixon BE, editor. *Health information exchange: navigating and managing a network of health information systems.* 2nd ed. Amsterdam; Boston: Academic Press, an imprint of Elsevier; 2023.
16. Kraus EM, Brand B, Hohman KH, Baker EL. New directions in public health surveillance: using electronic health records to monitor chronic disease. *J Public Health Manage Pract.* 2022;28(2):203–6.
17. Centers for Disease Control. Behavioral Risk Factor Surveillance System Weighting BRFSS Data [Internet]. 2015 [cited 2023 Feb 14]. Available from: https://www.cdc.gov/brfss/annual_data/2015/pdf/weighting_the_data_webpage_content.pdf
18. Trafimow D. Confidence intervals, precision and confounding. *New Ideas Psychol.* 2018;50:48–53.
19. Iachan R, Pierannunzi C, Healey K, Greenlund KJ, Town M. National weighting of data from the behavioral risk factor surveillance system (BRFSS). *BMC Med Res Methodol.* 2016;16(1):155.
20. Tatem KS, Romo ML, McVeigh KH, Chan PY, Lurie-Moroni E, Thorpe LE, et al. Comparing prevalence estimates from Population-Based surveys to inform surveillance using electronic health records. *Prev Chronic Dis.* 2017;14:E44.
21. Allen KS, Valvi N, Gibson PJ, McFarlane T, Dixon BE. Electronic health records for population health management: comparison of electronic health Record-Derived hypertension prevalence measures against established survey data. *Online J Public Health Inf.* 2024;16:e48300.
22. Lakens D. Equivalence tests: A practical primer for t tests, correlations, and Meta-Analyses. *Soc Psychol Personal Sci.* 2017;8(4):355–62.
23. Beaglehole R, Ebrahim S, Reddy S, Voûte J, Leeder S, Chronic Disease Action Group. Prevention of chronic diseases: a call to action. *Lancet.* 2007;370(9605):2152–7.
24. Dixon B, Staes C, Acharya J, Allen K, Hartsell J, Cullen T et al. Enhancing the nation's public health information infrastructure: a report from the ACMI symposium. *Journal of the American Medical Informatics Association.* 2023;14.
25. Acharya J, Staes C, Allen K, Hartsell J, Cullen T, Lenert L et al. Strengths, Weaknesses, Opportunities, and Threats for the Nation's Public Health Information Infrastructure: Synthesis of Discussion from the 2022 ACMI Symposium. Under Review.
26. Cox C, Rae M, Kates J, Wager E, Ortaliza J, Dawson L. A Look at Federal Health Data Taken Offline. *KFF Policy Watch* [Internet]. 2025 Feb 2 [cited 2025 Mar 2]; Available from: <https://www.kff.org/policy-watch/a-look-at-federal-health-data-taken-offline/>
27. Schneider KL, Clark MA, Rakowski W, Lapane KL. Evaluating the impact of non-response bias in the behavioral risk factor surveillance system (BRFSS). *J Epidemiol Community Health.* 2012;66(4):290–5.
28. Comer KF, Gibson PJ, Zou J, Rosenman M, Dixon BE. Electronic health record (EHR)-Based community health measures: an exploratory assessment of perceived usefulness by local health departments. *BMC Public Health.* 2018;18(1):647.
29. Institute of Medicine (U.S.) C on a NSS for C and SCD. A Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung Diseases. [Internet]. National Academies Press (US). 2011. (Existing Surveillance Data Sources and Systems). Available from: <https://www.ncbi.nlm.nih.gov/books/NBK83157/>
30. Dixon BE, Zou JF, Comer KF, Rosenman M, Craig JL, Gibson P. Using Electronic Health Record Data to Improve Community Health Assessment. 2016 [cited 2023 Dec 29]; Available from: <http://uknowledge.uky.edu/frontiersinphss/vol5/iss5/8/>
31. Stull DE, Leidy NK, Parasuraman B, Chassany O. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. *Curr Med Res Opin.* 2009;25(4):929–42.
32. He S, Park S, Fujii Y, Pierce SL, Kraus EM, Wall HK, et al. State-Level hypertension prevalence and control among adults in the U.S. *Am J Prev Med.* 2024;66(1):46–54.
33. Johnson KJ, Goss CW, Thompson JJ, Trolard AM, Maricque BB, Anwuri V, et al. Assessment of the impact of the COVID-19 pandemic on health services use. *Public Health Pract (Oxf).* 2022;3:100254.
34. Taber JM, Leyva B, Persoskie A. Why do people avoid medical care? A qualitative study using National data. *J Gen Intern Med.* 2015;30(3):290–7.
35. Lim MT, Lim YMF, Tong SF, Sivasampu S. Age, sex and primary care setting differences in patients' perception of community healthcare seeking behaviour towards health services. *Tu WJ, editor. PLoS ONE.* 2019;14(10):e0224260.
36. Hing E, Cherry DK, Woodwell DA. National ambulatory medical care survey: 2004 summary. *Adv Data.* 2006;(374):1–33.
37. HealthIT.gov. Trusted Exchange Framework and Common Agreement (TEFCA) [Internet]. [cited 2023 Feb 14]. Available from: <https://www.healthit.gov/topic/interoperability/policy/trusted-exchange-framework-and-common-agreement-tefca>
38. Tripathi M, Yeager M. TEFCA Live! The Future of Network Interoperability is Here [Internet]. *Health Affairs Forefront.* 2023 [cited 2023 Dec 29]. Available from: <https://www.healthaffairs.org/content/forefront/tefca-live-future-network-interoperability-here>
39. Angier H, Gold R, Gallia C, Casciato A, Tillotson CJ, Marino M, et al. Variation in outcomes of quality measurement by data source. *Pediatrics.* 2014;133(6):e1676–1682.
40. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: clinical versus statistical significance. *Perspect Clin Res.* 2015;6(3):169–70.
41. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and impact of Real-World clinical data for the practicing clinician. *Adv Ther.* 2018;35(11):1763–74.
42. Hasnain-Wynia R, Van Dyke K, Youdelman M, Krautkramer C, Ivey SL, Gilchick R, et al. Barriers to collecting patient race, ethnicity, and primary Language data in physician practices: an exploratory study. *J Natl Med Assoc.* 2010;102(9):769–75.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.