



HHS Public Access

Author manuscript

Chronic Illn. Author manuscript; available in PMC 2024 February 21.

Published in final edited form as:

Chronic Illn. 2023 September ; 19(3): 675–680. doi:10.1177/17423953221099043.

Using latent class analysis to inform the design of an EHR-based national chronic disease surveillance model

Laura Nasuti¹, Bonnie Andrews¹, Wenjun Li², Jennifer Wiltz³, Katherine H Hohman¹
[Guarantor], Miriam Patanian¹

¹National Association of Chronic Disease Directors, Decatur, USA

²Department of Public Health, Zuckerberg College of Health Sciences, University of Massachusetts Lowell, Lowell, USA

³Office of the Director, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, USA

Abstract

The Multi-state EHR-based Network for Disease Surveillance (MENDS) developed a pilot electronic health record (EHR) surveillance system capable of providing national chronic disease estimates. To strategically engage partner sites, MENDS conducted a latent class analysis (LCA) and grouped states by similarities in socioeconomic, demographics, chronic disease and behavioral risk factor prevalence, health outcomes, and health insurance coverage. Three latent classes of states were identified, which inform the recruitment of additional partner sites in conjunction with additional factors (e.g. partner site capacity and data availability, information technology infrastructure). This methodology can be used to inform other public health surveillance modernization efforts that leverage timely EHR data to address gaps, use existing technology, and advance surveillance.

Keywords

Chronic disease; surveillance; electronic health record; health information technology

Corresponding author: Miriam Patanian, National Association of Chronic Disease Directors, 325 Swanton Way, Decatur, GA 30030, USA. mpatanian_ic@chronicdisease.org.

Contributorship

LN was the primary author on this manuscript and analyzed output and sensitivity analyses for the paper. BA aggregated the data from several publicly available sources, ran the original model, and prepared an initial report of the latent class analysis that summarized early results. WL and a colleague conceived of the idea and worked with BA to design and implement the original iteration. JW provided critical review and revision of the manuscript throughout its development and administrative and technical support including leadership for CDC clearance. KHH conducted additional data iterations and a critical review of the manuscript for important alignment with the MENDS project. MP obtained funding for the project, as well as provided oversight on the implementation of this work and additional administrative and technical support. All authors contributed to and reviewed the final manuscript.

Declaration of conflicting interests

BA is a psychologist in a private practice and receives payment from insurance providers for her services. The other authors of this report have no conflicts of interest and no financial disclosures to report. Most authors have had contracts with the National Association of Chronic Disease Directors to conduct this work.

Introduction

In 2018, National Association of Chronic Disease Directors (NACDD), funded by Centers for Disease Control and Prevention (CDC), initiated the Multi-state EHR-based Network for Disease Surveillance (MENDS) to develop a near real-time, chronic disease surveillance system (<https://chronicdisease.org/page/MENDSINFO/>).¹ MENDS, modeled on a Massachusetts program,² aims to support chronic disease surveillance and inform public health program and policy development. An essential attribute of a surveillance system is that it is representative of the intended population to ensure accurate estimates of disease incidence, prevalence, quality, and other health events.³ Statistical methods that help group states, counties, or other geographies that support creating a representative population for key public health activities are needed. This short report explores the use of a latent class analysis (LCA) to inform state selection.

Methods

MENDS employed a latent class analysis (LCA) to group 50 states and the District of Columbia according to their similarities in socioeconomic, demographics, health insurance coverage, chronic disease and behavioral risk factor prevalence, based on prior methods developed for selection of representative comparison communities for a program evaluation.⁴ The LCA examined states, as opposed to communities or counties, as the organizing geographic unit for two reasons: 1) partner sites (e.g. health systems, health information exchanges) may not be limited to a smaller geography such as county, and 2) some information needed for this modeling was not easily available at smaller geographic levels. The goal of this LCA was to support data-informed selection of additional partner sites beyond the initial four sites identified for the project.

Data used for the LCA were from four broad domains: sociodemographic, prevalence of health-related risk factors and chronic disease, chronic disease-related health outcomes (e.g. heart disease deaths), and health insurance status. All data for the study were from publicly available sources (noted in Table 1): U.S. Census population estimates, the U.S. Census American Community Survey, and CDC Chronic Disease Indicators. These data sets had no personally identifiable information. To preserve stability of the LCA model given the relatively small number of units examined ($n = 51$), only 12 indicators from the four domains were used in the analysis.³ An indicator on persons ever-diagnosed with depression was added due to its high prevalence and possible negative effects on chronic disease self-management.^{5,6} Each variable was dichotomized prior to analysis to capture relative differences among the states as in Jiang et al.⁷ Analysis was performed using the Proc LCA package for SAS version 9.4.^{8,9}

Results

Among alternative models, this three-class solution was identified as the best fitting model in 51% of iterations. This model had strong classification certainty (entropy = 0.97, minimum class probability = 0.78). In terms of model fit, the Akaike Information Criteria was 532.30, the Bayesian Information Criterion (BIC) was 623.10, and the sample size-

adjusted BIC was 475.54. Core clusters of states remained relatively stable through many different model iterations (not shown). While model fit statistics indicated that a four-class solution could also be a fit to the data (log likelihood = -397.66 compared to log likelihood = -418.29 for a 3-class model), the small sample size (50 states and District of Columbia) and replicability of different iterations of the model led to selecting the three-class solution.

The three identified classes have a clear geographic orientation and strong similarities within each class. Class 1, the smallest class, contains 14 states and the District of Columbia (Figure 1). The states in this class have younger populations, lower proportions of non-Hispanic White residents, higher proportions of Hispanic and non-Hispanic Asian residents, and lower rates of risk factors such as smoking, obesity, and depression (means for each class shown in Table 1; item response probabilities not shown). Class 2, the largest class, comprises 19 states that have higher percentages of non-Hispanic Black residents, lower percentages of people reporting that they have health insurance, higher percentages of people living below poverty level, and higher rates of risk factors for chronic disease. Class 3 has 17 states including many of the Great Plains states and less populated states in the Northeast. These states have higher proportions of non-Hispanic White residents and lower proportions of non-Hispanic Black residents. The states within each class do vary in population size (e.g. California and DC are in the same class), as evidenced by the large standard deviation for total population. Detailed characteristics of states and the District of Columbia by latent class were also compiled and are available upon request.

Discussion

The LCA analysis identified three distinct groups of states, which aids in MENDS project planning and may serve as an approach for similar efforts with chronic disease surveillance to consider. States within each class exhibited strong similarities in domains (e.g. sociodemographic) while these differed notably among the classes. In addition, the three classes have a clear geographic orientation. These findings suggest that the three classes are cohesive and can be used to identify states when selecting additional sites. The different distribution of characteristics in the three classes warrants additional investigation to understand the underlying factors driving these patterns (e.g. population chronic disease burden, insurance status, public health infrastructure, etc. The National Health and Nutrition Examination Survey (NHANES) used a similar strategy to group states and territories into four strata according to similarities in sociodemographics and health, but not based on LCAs.¹⁰ Although our LCA is limited by the small number of units under analysis, it generated results to inform continued recruitment of sites. Similar LCAs have been used to classify cities/towns into categories of chronic disease risk despite small sample sizes.⁷

Use of EHR data has the potential to advance public health surveillance and guide public health interventions,¹¹ which may include approaches at various geographic levels. MENDS presents one opportunity to modernize chronic disease surveillance methods with timely EHR data. These results are only one part of the planning process, especially for EHR surveillance, which relies on engagement of clinical partner sites and resources for implementation. In the MENDS pilot, decisions to identify four initial partner sites were made prior to conducting the LCA analysis and relied upon partners' ability and

interest to engage in the pilot. For future expansion, the underlying data from states in under-represented latent classes will help inform the selection of partner sites to increase the representativeness of MENDS. Although this work was undertaken as part of surveillance planning at a national level, it could also be replicated at smaller geographies (e.g. county) for other surveillance planning purposes.

Acknowledgements

We acknowledge the following people for their contributions into the development of this publication: Amanda Martinez, MPH, MSN, RN; Kathy Foell, MS; Kayla Craddock, MPH; Jeanne Alongi, DrPH, MPH; and Marti Macchi, MEd, MPH, National Association of Chronic Disease Directors. Ms Martinez provided writing assistance and supported the preparation of submission materials. In addition, we wish to acknowledge Thomas G. Land, PhD, former University of Massachusetts Medical School, who contributed to the idea and its original iteration, and Matthew Ritchey, PT, DPT, OCS, MPH and Adam Vaughan, PhD, MPH, MS, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division for Heart Disease and Stroke Prevention, who provided input that was adapted into this research brief. The map was created with free map software at mapchart.net. No copyrighted materials were used in this article.

The creation of this information is associated with Multi-state EHR-based Network for Disease Surveillance (MENDS) (<https://chronicdisease.org/page/MENDSINFO/>). The MENDS network leverages three software applications: Electronic medical record Support for Public Health (ESP), PopMedNet, and RiskScape. ESP (<https://www.esphealth.org/>) is an open-source software application that extracts electronic health data, organizes the data into a standard format stored across multiple data tables, and applies algorithms to identify conditions of public health interest. PopMedNet (<https://www.popmednet.org/>) is a software application that allows for the querying of the ESP data tables. RiskScape (<https://www.esphealth.org/riskscape>) is a software application that supports sense-making and action by providing summaries and visualizations of the ESP data. MENDS is the national implementation of these three applications.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, as part of a financial assistance award totaling \$1,890,000, 100 percent funded by CDC/HHS (grant number #5NU38OT000286). The findings and conclusions in this report are those of the authors and do not necessarily represent the official views of, nor an endorsement by, CDC/HHS, or the U.S. Government.

References

1. National Association of Chronic Disease Directors. Multi-state EHR-based national disease surveillance (MENDS), <https://chronicdisease.org/page/MENDSINFO/>. (2019, accessed 1 November 2021).
2. Vogel J, Brown JS, Land T, et al. MDPHnet: secure, distributed sharing of electronic health record data for public health surveillance, evaluation, and planning. *Am J Public Health* 2014; 104: 2265–2270. [PubMed: 25322301]
3. German RR, Lee LM, Horan JM, et al. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. *MMWR Recomm Rep*. 2001; 50: 1–35, <https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm>. (2001, accessed 11 March 2022).
4. Arcaya M, Reardon T, Vogel J, et al. Tailoring community-based wellness initiatives with latent class analysis – Massachusetts Community Transformation Grant Projects. *Prev Chronic Dis* 2014; 11: 130215.
5. Lin EHB, Katon W, Von Korff M, et al. Relationship of depression and diabetes selfcare, medication adherence, and preventive care. *Diabetes Care* 2004; 27: 2154–2160. [PubMed: 15333477]
6. Katon WJ, Lin EHB, Von Korff M, et al. Collaborative care for patients with depression and chronic illnesses. *N England J Med* 2010; 363: 2611–2620. [PubMed: 21190455]
7. Jiang Y, Dumont DM, Cooper T, et al. A latent class model to identify city/town chronic disease patterns. *Prev Med* 2015; 73: 139–144. [PubMed: 25602912]

8. Lanza ST, Dziak JJ, Huang L, et al. PROC LCA & PROC LTA users' guide (Version 1.3.2). University Park: The Methodology Center, Penn State, 2015, https://www.methodology.psu.edu/files/2019/03/proc_lca_lta_1-3-2-1_users_guide-2ggq4d3.pdf (2015, accessed 1 November 2021).
9. SAS® [Software]. Version 9.4. Cary, NC: SAS Institute, 2014.
10. Chen TC, Clark J, Riddles MK, et al. National health and nutrition examination survey, 2015–2018: sample design and estimation procedures. National Center for Health Statistics. Vital Health Stat 2020; 2: 1–35.
11. Birkhead GS, Klompas M, and Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health* 2015; 36: 345–359. [PubMed: 25581157]

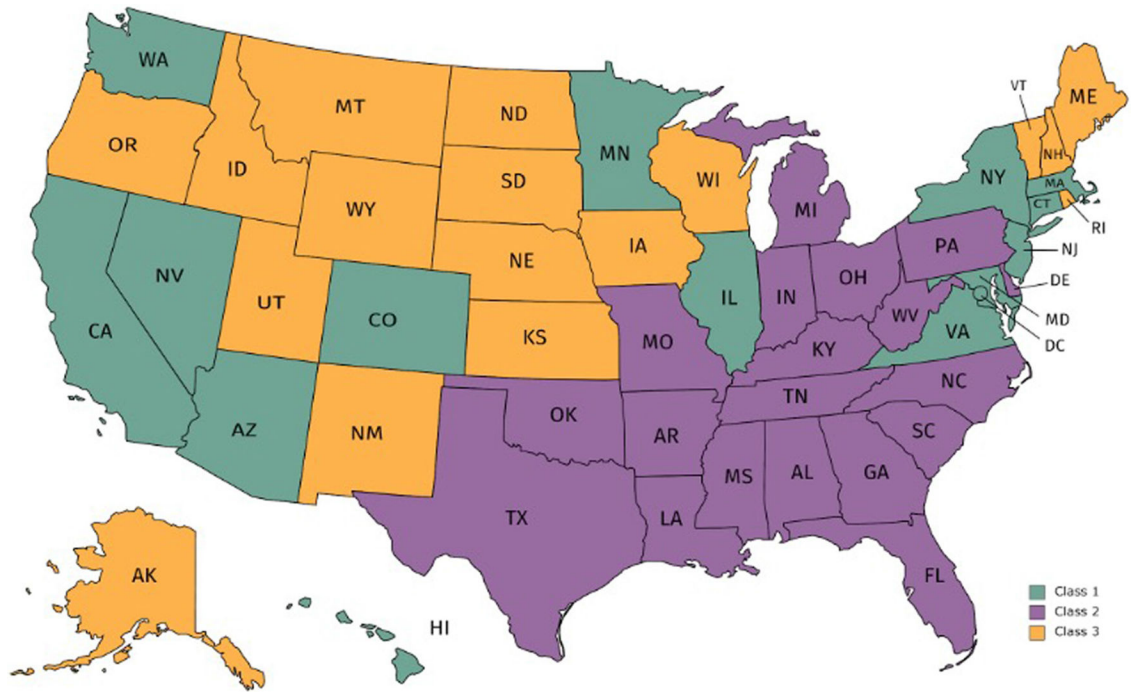


Figure 1.
States and district of Columbia by latent class.

Table 1.

State characteristics by latent class and overall national mean.

Indicator	Class 1 (n = 15)	Class 2 (n = 19)	Class 3 (n = 17)	National (n = 51)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Total population ^a	9,130,359 (9,603,693)	8,252,728 (6,911,308)	1,965,306 (1,453,934)	6,415,048 (7,343,307)
Percent of population by race/ethnicity (%) ^b				
White non-Hispanic	55.7 (15.2)	68.0 (12.6)	78.6 (13.3)	67.9 (16.2)
Black non-Hispanic	12.6 (11.5)	17.6 (9.5)	2.8 (1.9)	11.2 (10.5)
Hispanic	17.9 (9.4)	9.0 (9.0)	10.5 (10.9)	12.1 (10.3)
Asian non-Hispanic	8.9 (8.1)	2.5 (1.1)	2.5 (1.4)	4.4 (5.3)
Percent of population 65 and older ^b	14.6 (1.3)	15.5 (1.8)	15.2 (2.2)	15.2 (1.8)
Percent of population below poverty ^b	12.6 (2.6)	16.4 (2.2)	12.7 (2.7)	14.1 (3.0)
Percent of adults with obesity ^c	27.0 (2.8)	33.8 (2.4)	30.2 (3.0)	30.6 (3.9)
Percent of adults reporting current smoking ^c	14.3 (1.6)	20.3 (2.9)	16.5 (2.6)	17.3 (3.5)
Percent of adults reporting they had ever been diagnosed with diabetes ^c	9.8 (1.3)	12.3 (1.4)	9.2 (1.2)	10.5 (1.9)
Percent of adults reporting they had ever been told they have high blood pressure ^c	30.0 (2.4)	36.8 (3.5)	30.6 (2.1)	32.7 (4.2)
Age-standardized rate of death from stroke per 100,000 population ^c	43.0 (5.8)	53.2 (4.0)	42.1 (5.4)	46.5 (7.2)
Age-standardized rate of coronary heart disease deaths per 100,000 population ^c	109.1 (21.9)	131.9 (23.9)	110.2 (19.1)	118.0 (24.0)
Percent of adults reporting they had ever been told they have a form of depression ^b	17.2 (2.6)	21.6 (2.8)	21.5 (2.3)	20.3 (3.2)
Percent of adults ages 18-64 reporting they do not have health insurance ^b	11.5 (3.5)	15.6 (5.3)	12.4 (3.3)	13.3 (4.5)

n = total number of states [including District of Columbia] in each class.

^aU.S. Census Bureau. 2017 United States Census Population Estimates. Available at: <https://data.census.gov/cedsci/>.

^bU.S. Census Bureau. American Community Survey 5-year Estimates, 2013-2017. Available at: <https://data.census.gov/cedsci/>.

^cCenters for Disease Control and Prevention, Chronic Disease Indicators, 2017. Available at: <https://www.cdc.gov/cdi/>.