Moving Beyond Simple Risk Prediction: Segmenting Patient Populations Using Consumer Data

Mandana Rezaeiahari *

Department of Health Policy and Management, University of Arkansas for Medical Sciences, Little Rock, AR, United States

INTRODUCTION

There has been growing interest among health systems in population health (1, 2). Population health aims to improve the overall health of a population across the full continuum of care by more targeted, effective and coordinated health services (3). Given the rising trend of aging population and chronic disease burden, managing population health becomes more important for health systems trying to control cost (4).

In order to improve outcomes and efficiency, health systems need to customize care and interventions based on identified risks and costs (5). One of the systematic approaches in the literature for targeting interventions to subgroups of patients with different needs is population segmentation or risk-stratification (referred as *patient segmentation* in the remainder of this paper). Population segmentation that divides a population into groups with related service needs is an important foundation for effective and sustainable care delivery (6–8). Segmentation divides patients into distinct groups with specific needs, characteristics or behaviors and allows for health services to be organized around patients with similar needs (7). Patient segmentation models are becoming essential element of healthcare management due to the increase in the number of programs that incentivize value-based care (9).

Although patient segmentation models can help design interventions targeting subgroups of patients, they are often based on International Classification of Diseases (ICD) codes found in electronic health records (EHRs) and/or insurance claims data and lack important social risk factors that are essential for designing interventions. World Health Organization defines social determinants of health (SDOH) as the conditions in which people are born, grow, work, live, and age (10). These factors include economic policies and systems, development agendas, social norms, social policies and political systems (10). There are numerous studies demonstrating social factors acting as powerful determinants on multiple health outcomes including coronary heart disease (11), breast cancer (12), childhood obesity (13) and end-stage renal failure (14). Literature suggests that high utilizers of healthcare resources among Medicaid and uninsured population often have multiple chronic conditions (15, 16) and programs targeting this population collectively argue that social risk factors including but not limited to language, health literacy, unemployment, substance abuse and housing are important drivers of healthcare utilization (17, 18).

Most of the current patient segmentation models use administrative billing data because insurance claims data provides a nearly complete view of patients' interactions with health care delivery system; therefore, it is a reliable source to extract utilization outcomes (19). Majority of the EHRs on the other hand contain data from clinical encounters occurring between individuals and providers within a single health system and hence miss out of network events (19). On the positive side of EHRs is that they offer more extensive data including family history, lab results, vital signs and symptoms which could help improve the population segmentation model (20). One

drawback of reliance on insurance claims data and EHRs is that they miss social and behavioral factors that complicate care (21). Although, there is a subset of ICD-10-CM codes, the Z codes, for documenting SDOH in EHRs, these codes are underutilized (22, 23). As such, SDOH Z codes may not reflect the actual burden of social needs experienced by patients. To address this gap, this paper presents the complementary benefit of consumer data when it is linked to EHRs or insurance claims data. The consumer marketing data include individual-level SDOH (including income, education, lifestyle variables, language spoken, household size, smoking status, life events, shopping activity) that are not available in the insurance claims data or majority of EHR data. The combined data provides 360-degree view of patients and can help predict the risk of repeat emergency room visits or hospital admissions (24). Inclusion of SDOH is essential to improve population health as medical interventions without addressing social determinants are not sustainable and effective. This unprecedented view into the lives of patients has significant potential to improve upon segmentation approaches relying exclusively on health plan or EHR data that lack measures or even decent proxies for fitness, diet and other SDOH which can profoundly alter the course of chronic diseases. A number of commercial companies provide marketing data that is well-utilized by organizations that subscribe to their services. Experian's ConsumerView[SM] U.S. database is one of the world's largest consumer database on more than 300 million individuals and 126 million households (25). ConsumerView[SM] U.S. database is compiled from hundreds of resources. For example, property and mortgage data are compiled from public records and county deeds while lifestyle and interest data are compiled from consumers who have completed self-reported surveys (25). Marketing companies match and mange patient identity across the healthcare ecosystems enabling the linkage of datasets across channels and silos (26, 27). According to Acxiom, two-thirds of hospitals actively use or want third-party consumer and lifestyle data to improve patient care (24).

CURRENT PATIENT SEGMENTATION MODELS

There are two major approaches for conducting population segmentation in the literature. Expert-driven approaches are informed by expert consensus while data-driven approaches use statistical analysis such as clustering to segment a population (28). John Hopkins Adjusted Clinical Group (ACG) system and the Clinical Risk Group (CRG) system by 3M Health Information Systems are examples of expert-driven approaches (29, 30). The ACG system assigns each diagnosis code to one or more of 32 diagnosis groups referred to as Aggregated Diagnosis Groups (ADG). Both ACG and CRG system use diagnostic codes to classify patients into over 200 mutually exclusive risk groups (28). ADGs are assigned based on five features of conditions: duration, severity, diagnostic certainty, type of etiology and expected need for specialty care. The 3M CRG system assigns an individual five-digit classification code with first digit representing the core health status group, second through the fourth digit representing the base 3M CRG and the fifth digit identifying the severity-of-illness level (30). One drawback of expert-driven approaches is that they subjectively segment populations and no specific standards are set to derive the number of segments. Data-driven approaches generate evidence-based insights of population health status based on patient healthcare data to support policy decisions (1). There have been multiple studies using data-driven approaches to segment populations (19, 31, 32). Zhang et al. (33) developed a patient taxonomy with ten categories to divide high-cost Medicare Fee-For-Service patients. They found high-cost patients were most likely to have multiple chronic conditions, serious mental illness, serious medical illness and frailty (33). Low et al. (28) used cluster analysis and healthcare utilization data from electronic medical records to develop five segments of population (28). Concurrent with patient segmentation models developed by researchers, many predictive models based on SDOH have been developed by health payers and analytics companies. Most often these models are proprietary hence not available for review and scrutiny (34). For instance, a non-profit health insurance company used consumer data to develop a segmentation model to make informed adjustments to its Medicare marketing efforts (35).

DISCUSSION

As medical care is only responsible for 15 to 20% of preventable mortality in the US (36) and due to the increasing impact of social factors on health, it is now time to leverage data analytics to start to understand SDOH and its impact on health and design more social centered care coordination interventions (37). A recent critical review of patient segmentation models shows a lack of comprehensive models that integrates data from multiple sources, with a majority of the models limited to administrative billing data alone (21).

Healthcare organizations and payers should strive to link their traditional resources including EHRs and insurance claims data to consumer marketing data. Through this linkage, they can then apply advanced analytics to get tangible results that can be acted upon to improve quality of care and health outcomes. Specifically, more data driven approaches are needed to utilize available data to assess whether distinct patient subgroups might exist within population. For example, cluster analysis may be used to determine if individual level SDOH (based on consumer marketing data) and insurance claims, together can represent social, medical and behavioral health conditions to form specific relevant subgroup of patients. The proposed patient segmentation framework will facilitate healthcare resource planning and development of interventions to improve the healthcare delivery for each segment. This approach to segmentation will demonstrate heterogeneity in population groups with respect to age, morbidity, lifestyle, setting in which care was mostly used, etc. Therefore, depending on the patterns of utilization of care, complexity level of patients and lifestyle segmentation, various models of care will be needed. For instance, for "young or middle age and healthy" segment that focus little on preventive care and are fans of fast food, the

most important approaches may be disease prevention, health education and robust primary care, working with non-healthcare partners such as employers, community-based disease education in order to maintain the health status and promote healthy behavior. Patients with stable but chronic condition that are more interested in adopting technology, can instead benefit more from supportive self-management such as home-based self-monitoring tools to promote health empowerment. Patients with complex chronic conditions that are not managed well and live in neighborhoods with low levels of food access may require more multidisciplinary medical and social care coordination.

Some other examples of the opportunities as a result of linking consumer data to insurance claims and/or EHRs (not limited to patient segmentation) include reduction of obesity through increasing the relevance and effectiveness of weight loss engagement strategies by using consumer lifestyle segmentation variables including diet attitudes and motivations, gauging the receptivity of patients to different outreach channels (automated voice, live agent calls and text messages) using the digital media preference, age and education level, identifying food insecure households/individuals using frequency and dollars spend in food category particularly for individuals living in low-income and low food-access neighborhoods.

Unlike other traditional SDOH data sources that are only available at the county and/or zip code level, such as Area Health Resource Files (38), US Census County Business Patterns (39), and County Health Rankings (40), consumer data is available at the individual or household level. County level social data, although useful, only represent a profile of the community and does not reliably represent the profile of the individual patient. For example, research has shown that poverty is strongly associated with an increase in risk of dying, but simply living in a high-poverty area is not (41).

Despite the important opportunity that the consumer marketing data brings to healthcare, major concerns still exist about privacy of consumers. Linking consumer marketing data to EHRs and/or insurance claims data may increase informational risk (i.e., HIPAA violations), if strict data deidentification standards are not in place and/or data protections are applied inconsistently across various entities which collect, share and use the data (42). As such, any use cases of consumer data must be HIPAA compliant to ensure protection of "individually identifiable health information" (i.e., protected health information) (43). Some of the best practices to ensure compliance are safe sourcing (working with the source compilers of consumer data to ensure compliance), safe storage (reviewing and updating data privacy policies to control access), appropriate/ethical use of data (marketing data should never be used to deny access to anyone or result in health disparities) (44).

Other challenges of using consumer data include reproducibility and analytical challenges. Predictive models developed by the private sector are not shared publicly, therefore cannot be replicated by other researchers to ensure accuracy, validity and potential model bias (34). Additionally, researchers should be cautious when selecting the analytical approaches when it comes to the inclusion of marketing data to predict health outcomes. Highly flexible machine learning algorithms may select features (e.g., reality TV show from consumer interest data) to predict mortality which may not be clinically reasonable.

Despite the challenges discussed above, consumer marketing data may open up opportunities to health researchers to understand how individual level SDOH manifest throughout a person's life. Future patient segmentation models that incorporate SDOH from consumer marketing data have the potential to improve health and reduce health disparities by ensuring that the right patients will be intervened at the right time.

AUTHOR CONTRIBUTIONS

MR devised the idea, performed literature review, and wrote the manuscript.

REFERENCES

1. Vuik SI, Mayer E, Darzi A. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. *Popul Health Metr.* (2016). 14:44. doi: 10.1186/s12963-016-0115-z

2. Yan S, Kwan YH, Tan CS, Thumboo J, Low LL. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Med Res Methodol.* (2018) 18:121. doi: 10.1186/s12874-018-0584-9

3. Felt-Lisk, S, Higgins T. *Exploring the Promise of Population Health Management Programs to Improve Health.* Washington, DC: Mathematica Policy Research (2011).

4. Nnoaham KE, Cann KF. Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? *BMC Public Health.* (2020) 20:798. doi: 10.21203/rs.2.122 72/v2

5. *Value Transformation Framework Action Guide.* (2019). Available online at: https://www.nachc.org/wp-content/uploads/2019/03/Risk-Stratification-Action-Guide-Mar-2019.pdf (accessed May 28, 2021).

6. Chong JL, Matchar DB. Benefits of population segmentation analysis for developing health policy to promote patient-centred care. *Ann Acad Med Singap.* (2017) 46:287–9.

7. Vuik SI, Mayer EK, Darzi A. Patient segmentation analysis offers significant benefits for integrated care and support. *Health Aff.* (2016) 35:769–75. doi: 10.1377/hlthaff.2015.1311

8. Lynn J, Straube BM, Bell KM, Jencks SF, Kambic RT. Using population segmentation to provide better health care for all: the "bridges to health" model. *Milbank Q.* (2007) 85:185–208. doi: 10.1111/j.1468-0009.2007.00483.x

9. *Early Adopters of the Accountable Care Model: A Field Report on Improvements in Health Care Delivery. Commonwealth Fund.* Available online at: https://www.commonwealthfund.org/publications/fund-reports/2013/mar/early-adopters-accountable-care-model-field-report-improvements (accessed May 18, 2021).

10. *Social determinants of health.* Available online at: https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1 (accessed June 21, 2021).

11. Kim D. The associations between US state and local social spending, income inequality, and individual all-cause and cause-specific mortality: the National Longitudinal Mortality Study. *Prev Med.* (2016) 84:62–8. doi: 10.1016/j.ypmed.2015.11.013

12. Shariff-Marco S, Yang J, John EM, Kurian AW, Cheng I, Leung R, et al. Intersection of race/ethnicity and socioeconomic status in mortality after breast cancer. *J Commun Health.* (2015) 40:1287–99. doi: 10.1007/s10900-015-0052-y

13. Flood TL, Zhao YQ, Tomayko EJ, Tandias A, Carrel AL, Hanrahan LP. Electronic health records and community health surveillance of childhood obesity. *Am J Prev Med.* (2015) 48:234–40. doi: 10.1016/j.amepre.2014.10.020

14. Hill KE, Gleadle JM, Pulvirenti M, McNaughton DA. The social determinants of health for people with type 1 diabetes that progress to end-stage renal disease. *Health Expect.* (2015) 18:2513–21. doi: 10.1111/hex.12220

15. Johnson TL, Rinehart DJ, Durfee J, Brewer D, Batal H, Blum J, et al. For many patients who use large amounts of health care services, the need is intense yet temporary. *Health Aff.* (2015) 34:1312–9. doi: 10.1377/hlthaff.2014.1186

16. *Faces of Medicaid III: Refining the Portrait of People with Multiple Chronic Conditions—Center for Health Care Strategies.* Available online at: https://www.chcs.org/resource/the-faces-of-medicaid-iii-refining-the-portrait-of-people-with-multiple-chronic-conditions/ (accessed May 14, 2021).

17. *Caring for High-Need, High-Cost Patients: What Makes for a Successful Care Management Program? Commonwealth Fund.* Available online at: https://www.commonwealthfund.org/publications/issue-briefs/2014/aug/caring-high-need-high-cost-patients-what-makes-successful-care (accessed May 14, 2021).

18. *Strategies to Reduce Costs and Improve Care for High-Utilizing Medicaid Patients: Reflections on Pioneering Programs—Center for Health Care Strategies.* Available online at: https://www.chcs.org/resource/strategies-to-reduce-costs-and-improve-care-for-high-utilizing-medicaid-patients-reflections-on-pioneering-programs/ (accessed May 14, 2021).

19. Kharrazi H, Chi W, Chang HY, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. *Med Care.* (2017) 55:789–96. doi: 10.1097/MLR.0000000000000754

20. Wilson J, Bock A. *The benefit of using both claims data and electronic medical record data in health care analysis.* (2012). Available online at: https://www.optum.com/content/dam/optum/resources/whitePapers/Benefits-of-using-both-claims-and-EMR-data-in-HC-analysis-WhitePaper-ACS.pdf (accessed June 22, 2021).

21. Jeffery AD, Hewner S, Pruinelli L, Lekan D, Lee M, Gao G, et al. Risk prediction and segmentation models used in the United States for assessing risk in whole populations: a critical literature review with implications for nurses' role in population health management. *JAMIA Open.* (2019) 2:205–14. doi: 10.1093/jamiaopen/ooy053

22. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016–2017. *Med Care.* (2020) 58:1037–43. doi: 10.1097/MLR.0000000000001418

23. Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International classification of diseases, tenth revision, clinical modification social determinants of health codes are poorly used in electronic health records. (2020) 99:e23818. doi: 10.1097/MD.0000000000023818

24. Acxiom. *The Power of Consumer and Lifestyle Data Iin Healthcare.* Available online at: https://www.acxiom.com/resources/infographic-the-power-of-consumer-and-lifestyle-data-in-healthcare/ (accessed June 22, 2021).

25. Experian. *Experian audience lookbook.* Available online at: https://www.experian.com/content/dam/marketing/na/assets/ems/marketing-services/documents/product-sheets/audience-lookbook.pdf (accessed June 19, 2021).

26. *Healthcare Marketing—Predictive Analytics, Database Solutions, Strategy.* Available online at: https://www.acxiom.com/healthcare/ (accessed June 22, 2021).

27. ConsumerView SM. *Tap Into the Power of the World's Largest Consumer Database* (2018).

28. Low LL, Yan S, Kwan YH, Tan CS, Thumboo J. Assessing the validity of a data driven segmentation approach: a 4 year longitudinal study of healthcare utilization and mortality. *PLoS ONE.* (2018) 13:e0195243. doi: 10.1371/journal.pone.0195243

29. The Johns Hopkins ACG® System. *Excerpt from Version 11.0 Technical Reference Guide.* The Johns Hopkins ACG® System (2014).

30. *3MTM Clinical Risk Groups: Measuring risk, managing care.* (2016). Available online at: https://multimedia.3m.com/mws/media/765833O/3m-crgs-measuring-risk-managing-care-white-paper.pdf (accessed June 22, 2021).

31. Rinehart DJ, Oronce C, Durfee MJ, Ranby KW, Batal HA, Hanratty R, et al. Identifying subgroups of adult superutilizers in an urban safety-net system using latent class analysis. *Med Care.* (2018). 56:e1–9. doi: 10.1097/MLR.0000000000000628

32. Murphy SME, Castro HK, Sylvia M. Predictive modeling in practice : improving the participant identification process for care management programs using condition-specific cut points. *Popul Health Manag.* (2011) 14:205–10. doi: 10.1089/pop.2010.0005

33. Zhang Y, Grinspan Z, Khullar D, Unruh MA, Shenkman E, Cohen A, et al. Developing an actionable patient taxonomy to understand and characterize high-cost Medicare patients. *Healthcare.* (2020) 8:100406. doi: 10.1016/j.hjdsi.2019.100406

34. Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, et al. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. *JMIR Med Inform.* (2020) 8:e18084. doi: 10.2196/18084

35. Simpson M, Genovese A. *Leveraging consumer data to grow medicare market share.* (2016). Available online at: https://info.carrothealth.com/hubfs/Brochures%20and%20Whitepapers/Carrot%20Health%20-%20Leveraging%20Consumer%20Data%20to%20Grow%20Medicare%20Market%20Share.pdf?__hstc=122733652.515b5d9ff7a33417378b5a218fdca83f.1567383407805.1567386846670.1567398132639.3&__hssc=122733652.1.1567398132639 (accessed May 24, 2021).

36. McGinnis JM, Williams-Russo P, Knickman JR. The case for more active policy attention to health promotion. *Health Aff.* (2002) 21:78–93. doi: 10.1377/hlthaff.21.2.78

37. Mackenbach JP. The contribution of medical care to mortality decline: mcKeown revisited. *J Clin Epidemiol.* (1996) 49:1207–13. doi: 10.1016/S0895-4356(96)00200-4

38. *Area Health Resources Files.* Available online at: https://data.hrsa.gov/topics/health-workforce/ahrf (accessed June 22, 2021).

39. Bureau UC. *County Business Patterns (CBP).* Available online at: https://www.census.gov/programs-surveys/cbp.html (accessed June 22, 2021).

40. *County Health Rankings & Roadmaps.* Available online at: https://www.countyhealthrankings.org/ (accessed June 22, 2021).

41. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. *PLoS Med.* (2010) 7:e1000316. doi: 10.4016/19865.01

42. Rahimzadeh V. A policy and practice review of consumer protections and their application to hospital-sourced data aggregation and analytics by third-party companies. *Front Big Data.* (2021) 3:44. doi: 10.3389/fdata.2020.603044

43. *Enhance Healthcare Analytics with Consumer Data.* (2019). Available online at: https://marketing.acxiom.com/US-Enhance-Healthcare-eb-main2.html?&utm_source=website&utm_medium=owned&utm_campaign=EnhancedHCeB (acessed October 3, 2021).

44. *The 3 keys to compliance for healthcare marketing data—Healthcare Blog.* Available online at: https://www.experian.com/blogs/healthcare/2019/05/the-3-keys-to-compliance-for-healthcare-marketing-data/ (accessed May 18, 2021).