Empiric Segmentation of High-risk Patients: A Structured Literature Review

Jonathan Arnold, MD, MS, MSE; Joshua Thorpe, PhD, MPH; Janke Mains-Mason, MPH; and Ann-Marie Rosland, MD, MS

ABSTRACT

OBJECTIVES: Empiric segmentation is a rapidly growing, learning health system approach that uses large health care system data sets to identify groups of high-risk patients who may benefit from similar interventions. We aimed to review studies that used data-driven approaches to segment high-risk patient populations and describe how their designs and findings can inform health care leaders who are interested in applying similar techniques to their patient populations.

STUDY DESIGN: Structured literature review.

METHODS: We searched for original research articles published since 2000 that identified high-risk adult patient populations and applied data-driven analyses to segment the population. Two reviewers independently extracted study population source and criteria for high-risk designation, segmentation method, data types included, model selection criteria, and model results from the identified studies.

RESULTS: Our search identified 224 articles, 12 of which met criteria for full review. Of these, 8 segmented high-risk patients and 4 segmented diagnoses without assigning patients to unique groups. Studies segmenting patients more often had clinically interpretable results. Common groups were defined by high prevalence of diabetes, cardiovascular disease, psychiatric conditions including substance use disorders, and neurologic disease (eg, stroke). Few studies incorporated patients' functional or social factors. Resulting patient and diagnosis clusters varied in ways closely linked to the model inputs, patient population inclusion criteria, and health care system context.

CONCLUSIONS: Empiric segmentation can yield clinically relevant groups of patients with complex medical needs. Segmentation results are context dependent, suggesting the need for careful design and interpretation of segmentation models to ensure that results can inform clinical care and program design in the target setting.

Am J Manag Care. 2022;28(2):e69-e77. doi:10.37765/ajmc.2022.88752

It is well documented that for any health care system patient population, there is a small subset of patients who have disproportionally high medical utilization, incur the majority of medical costs, and are at highest risk of poor health outcomes ("high-risk patients").[1-3] Despite the fact that health care systems and payers have become very successful at identifying who among their patients are high risk and high cost, interventions that apply one approach to all patients who are identified as high risk have not improved health outcomes.[4-8] The challenge for health care organizations remains this: how to translate their ability to predict risk into interventions that provide high-risk patients the care that they need to prevent unnecessary hospitalizations and improve health outcomes.[9]

Heterogenous health profiles and health care needs among high-risk patients make it difficult to design an effective one-size-fits-all intervention approach.[10] However, fully individualized care management for each patient takes more resources per patient than health care systems can typically provide. Similarly, addressing high-risk patients' health conditions one at a time is ineffective and inefficient,[11] as most high-risk patients have numerous concurrent health conditions and use health care across a variety of settings. Dividing high-risk patients into groups based on similar sets of health conditions and health care needs would allow health care systems to develop a set of interventions optimized for each group. For this reason, the National Academy of Medicine 2017 report *Effective Care for High-Need Patients* called for a "taxonomy that presents holistic guidance on how care and finite resources should be targeted and delivered to improve the health of high-need individuals."[5]

Following the lead of marketing and other industries,[12] health care systems are starting to use patient data to empirically segment their high-risk populations.[10,13] Previous attempts to segment high-risk populations were driven by descriptive, a priori assumptions about which conditions define patients with similar health care needs.[12,14,15] Those approaches may work well for general populations with low morbidity for whom comprehensive epidemiologic information is available. For populations of higher-risk patients, however, this

TAKEAWAY POINTS

Health care systems are increasingly using patient data to segment high-risk populations, but these analyses have not been widely translated into risk-reducing interventions. In this review of 12 published empiric high-risk patient segmentation studies, we find:

- Groups identified varied based on patient inclusion criteria, the patient factors that are input into models, data sources, and model type.
- Groups were often defined by common chronic health conditions, but few studies input patients' functional or social factors into models.
- We provide summary recommendations that health care systems can use to guide future high-risk patient segmentation analyses so that they can best inform design of group-tailored interventions.

TABLE 1. PubMed Search Query

Multimorbidity/high-risk/high-cost	Methods	Chronic disease
MeSH • Multimorbidity • Delivery of health care, integrated/economics • Health care costs • Health services needs and demand/economics Free text • Multimorbidity • High-cost, high-need • Complex(ity) primary care	MeSH • Cluster analysis • Machine learning • Models, statistical Free text • Latent class analysis • Item response theory • Mixture item response theory • Cluster analysis • Predictive modeling	MeSH • Chronic disease • Chronic disease/economics • Chronic disease/epidemiology Free text • Chronic disease
Final search		
((((((((("Multimorbidity"[MeSh]) OR (((("Delivery of Health Care, Integrated/economics"[MeSH]) OR "Health Care Costs"[MeSH]) OR ("Health Services Needs and Demand/economics"[MeSH]))))) AND (("Cluster Analysis"[MeSH]) OR "Models, Statistical"[MeSH])) AND (("Chronic Disease"[MeSH]) OR ("Chronic Disease/economics"[MeSH] OR "Chronic Disease/epidemiology"[MeSH]))) AND ("Adult"[MeSH]))))) OR (((((((multimorbidity) OR (((high-cost high-need) OR high-cost) OR high-need)) OR ((complex*) AND primary care))) AND ((((((((latent class analysis) OR latent class)) OR ((item response theory) OR mixture item response theory)) OR ((cluster analysis) OR clustering)) OR predictive modeling)) AND chronic disease)))		

MeSH, Medical Subject Headings.

approach is limited by assumptions about the similarities of high-risk populations and the ability of descriptive methods to distinguish combinations of only 2 or 3 conditions at a time.[16-18] Prespecified disease combinations with known high degrees of association in a general population, such as diabetes and heart disease, may also be common within a high-risk patient population, but these prespecified combinations may not best distinguish subgroups from one another or help identify drivers of health outcomes. In addition, simply describing which diagnoses co-occur can overlook patients with different but related sets of conditions that have "concordant" management approaches and can be grouped for similar clinical interventions.[19,20] Generic a priori classification schemes also do not allow health care systems to customize interventions to the segments that exist within their own high-risk population. Data-driven segmentation addresses these limitations by allowing patient data to "speak for itself" by using complex patterns in the large volume of health data generated by high-risk patients to reveal subgroups. Health care systems can then design and implement a limited set of interventions customized to these empiric groups and use outcomes data to feed back into continuous improvement of care for each group.[21]

In the last decade, empiric segmentation of high-risk patients has exploded in popularity. Several publications have outlined the reasons that segmentation may be useful to high-risk patient care[9,21] and others have outlined options for segmentation modeling methods.[22] However, there is not consensus on the best approach to using segmentation models with high-risk populations. There is now an opportunity to summarize patterns among methods and results from this first wave of applied studies to inform future segmentation studies and increase the likelihood that these analyses can improve patient care.

We conducted a structured review of published research articles that used empiric techniques to segment a health care system's high-risk patients or their health conditions into data-driven groups. We describe how each study approached model design and interpretation and discuss how these decisions may have influenced study results. We then summarize patterns across studies and address implications for how segmentation study design can be optimized to ensure that findings will be most applicable to care improvements for high-risk patients.

METHODS

We conducted a structured literature review of original research articles that identified medically high-risk health care system patient populations and applied data-driven analysis approaches to segment the population.

Search Strategy

We first used PubMed to search Medline and PubMed Central databases for original English-language research articles published between January 2000 and January 2020. We included relevant Medical Subject Headings and free-text terms as outlined in **Table 1**. We then examined bibliographies from each article identified by this initial search to identify additional potentially relevant studies.

Finally, we conducted additional hand searches using keywords and articles entered into Google Scholar.

Study Inclusion and Exclusion Criteria

A total of 219 articles resulted from our initial Medline search, and an additional 5 potential articles were identified via reference review and hand searching. We excluded "gray" literature such as conference abstracts, unpublished theses, or industry reports. Studies were included if their population of focus was adults (18 years or older) and selected based on (1) high observed health care costs or utilization, (2) high predicted health care costs or utilization, or (3) high predicted risk for poor health outcomes, including patients with documented multimorbidity. Although age itself may be considered a risk factor for poor health outcomes, we excluded study populations solely defined by older age. We excluded studies that segmented general populations not otherwise known to be high risk or high cost, as well as studies conducted among narrower disease- or syndrome-specific populations. Studies were included if they used empiric, data-driven modeling strategies to segment their population[23] (ie, they did not apply a priori segments). Included studies applied modeling strategies either to the entire high-risk population or to broad strata within that population (such as men/women). To determine whether studies from our search met the inclusion criteria listed above, 2 authors (J.A. and J.M.-M.) reviewed the abstract of each article, then the full text of any article in which inclusion information was not clear from the abstract. In one instance with lack of consensus, a third author (A.M.R.) reviewed the full article text to determine whether inclusion criteria were met.

Data Extraction

Two authors (J.A. and J.M.-M.) independently extracted prespecified data from the included articles and their published supplementary materials. A third author (A.M.R.) reviewed and resolved any conflicting results. For each publication, we extracted the source population, high-risk or high-cost selection criteria, segmentation method and reported model fit metrics, variables included in segmentation models, sources of variables (eg, electronic health record [EHR], patient report), the stated criteria for selecting the optimal segmentation result, and the resulting patient or diagnosis groups, as described by the original article authors.

RESULTS

We found 12 articles[24-35] that matched our inclusion criteria: 8 patient-clustering (**Table 2**[27-31,33-38]) and 4 diagnosis-clustering (**Table 3**[24-26,32]) articles. High-risk patients were selected based on high cost (1 study),[27] a risk-scoring system (3 studies),[28,29,35] and multimorbidity (8 studies).[24-26,30-34] High-risk study populations were drawn from the adult populations of the US Veterans Health Administration (2 studies),[24,28] US Medicare enrollees (1 study),[33] 1 US nongovernmental regional health system (Kaiser Permanente, 3 studies),[27,34,35] and non-US governmental (Switzerland, 1 study)[26] and regional (5 studies)[25,29-32] health systems. Most study populations were from 2010 or later, but one study included patients from 2009[30] and another from 1997 to 2000.[24] In 6 of 9 samples based in health care systems, patients were restricted to those receiving primary care in that health care system.[24-26,28,30,34]

Studies typically used between 20 and 80 patient data variables in their segmentation models, with a range from 12 to 263. Most studies limited clinical diagnosis inputs to those with a minimal population prevalence (ie, 1% of 5% of the study population), or applied judgment in selecting conditions most relevant to clinical interventions. The most common patient data inputs were clinical diagnoses, derived either from EHR or billing data. One study combined clinical diagnoses with utilization patterns also obtained from the EHR.[33] Three studies used patient-reported data on health status, functional status, or social needs.[33-35]

Of the 8 patient-clustering studies (Table 2[27-31,33-38]), 2 studies[30,31] used cluster analysis[39] and 5 used latent class analysis[40] or a closely related method.[27-29,33,35] One study used both and compared the results.[34] All 4 of the diagnosis-clustering studies (Table 3[24-26,32]) used cluster analysis[24-26,32] and 1 additionally used exploratory factor analysis.[41] Goodness-of-fit metrics were reported for 6 of the 14 models. No studies used patient outcomes to determine patient groupings (eg, "supervised" segmentation[42]).

With each empiric segmentation method, multiple solutions are possible, particularly in terms of the total number of resulting segments. Thirteen of the 14 analyses provided details on the criteria they used to select their "final" or "best" model solution, as outlined in Tables 2 and 3 under "Model Selection Method."[24,25,27-35] Six studies used only statistical or analytical model fit measures, without describing a role for clinical judgment in final model selection.[29-33,35] Four studies used statistical measures as an initial screen, then used clinical judgment for the final selection.[25,27,28,34] Clinical judgment as a criterion was rarely described beyond "clinical interpretability." However, 1 study described this in detail as including (1) whether the groups of conditions matched known epidemiology, (2) whether it would be clinically useful to be aware of co-occurring conditions within a cluster, and (3) whether the chronic conditions within the cluster would respond to similar clinical management approaches.[24] This was also the only study to use clinical judgment as the sole criteria for final model selection.

The segments derived from empiric modeling were highly influenced by the population, modeling approach, and variables chosen. Patient-clustering studies were more clinically interpretable than diagnosis-clustering studies and resulted in fewer segments (median, 6 vs 28). Similarly, studies that included clinical judgment in their model selection methods typically resulted in more clinically interpretable results compared with studies relying on purely statistical criteria. In the 2 studies that included functional limitations (eg, ability to perform activities of daily living) as inputs along with chronic condition diagnoses, these functional variables were key in differentiating resulting groups.[33,34]

TABLE 2. Patient-Clustering Studies From High-risk Populations[27-31,33-38]

Citation; N and source population	Study population year(s)	High-risk selection criterion	Segmentation method (model performance metrics)	Included variables	Model selection method	Resulting patient groups
High observed cost population						
Davis et al (2018)[27]; 21,183 US adults continuously enrolled in Kaiser Permanente Southern California	2010	Top 1% in total observed health care expenditures	LCA (average posterior probability)	53 acute and chronic condition diagnoses from EHR	BIC, clinical judgment	•Few comorbidities (33%) •CVD, CHF, COPD (17%) •Cancer, blood, autoimmune (14%) •ESRD, DM, metabolic (12%) •Sepsis, CVD, CHF, malnutrition (11%) •DM, CVD, CHF, retinopathy (8%) •CVA, head injury, paralysis/coma (5%)
High predicted utilization population						
Prenovost et al (2018)[28]; 68,400 US adults receiving primary care in the Veterans Health Administration	2014	10% sample of the top 10th percentile in predicted risk of hospitalization in the next year based on CAN-2-H score[36]	Mixture LCA-IRT models (goodness-of-fit metrics not reported)	31 chronic condition diagnoses from EHR	Information criteria (BIC) then clinical judgment of interpretability	•DM, CVD, CKD (35%) •Depression, bipolar, PTSD, DM (21%) •SUD, depression, PTSD, bipolar (15%) •Cancer, CVD, CKD (13%) •Liver (9%) •Cancer, depression (7%)
Buja et al (2018)[29]; 2691 Italian adults ≥65 years living in the Veneto region	After 2012 (end date not given)	High ACG-predicted health care utilization[37]	LCA (goodness-of-fit metrics not reported)	15 chronic condition diagnoses from EHR	BIC to select the optimal model	•Cancer, HTN, CHF (30%) •CVD, CHF, DM (22%) •Dementia, neurovascular, depression (19%) •Cancer (18%) •COPD, asthma, CHF (10%)
Rogers et al (2020)[35]; 2533 US adults enrolled in Kaiser Permanente	12/2015-11/2016	Predicted to be in the top 1% of health care utilizers using in-house predictive modeling[38]	LCA (goodness-of-fit metrics not reported)	14 patient-reported "social risks" obtained via phone survey[a]	BIC to select the optimal model	•High prevalence all SDOH, food insecurity (17%) •High-medium SDOH (13%) •Medium SDOH, caregiver assistance (17%) •Low SDOH (53%)
Multimorbidity population						
Keeney et al (2019)[33]; 4,156,594 US Medicare beneficiaries	2014	"High needs" based on multimorbidity, prior year utilization, and dependence in mobility/ADLs	LCA (smallest mean posterior probability, entropy)	12 clinical variables[b]	Log likelihood, BIC, and Lo-Mendell-Rubin likelihood ratio test	5-class model •≥6 comorbidities and hospitalizations (33%) •IHD with home care (23%) •IHD with hospitalization and SNF (22%) •Home care use (12%) •ADRD, NH use, and functional limitation (10%)

(*continued*)

In 2 of the 4 studies with patient samples identified by high cost or utilization, there was 1 resulting group defined by "few comorbidities"[27] or "low needs."[35] This "healthier" segment was the largest one in both of these studies (33% and 53% of patients, respectively). This implies that there were patient factors associated with cost or utilization that were not included in the models. Similarly, in 3 of the 4 studies that segmented patients with multimorbidity, the largest groups were defined as "healthy" (22%),[34] "nonspecific, common conditions" (34%-38%),[30] or "nonspecific, lower than typical prevalence of most diagnoses" (41%).[31] Patients in these groups typically had conditions that were common across all groups, such as hypertension, without additional conditions that distinguished their group from others.

Although the resulting groups varied significantly among studies, some patterns emerged. The most common groups were defined by high prevalence of diabetes, cardiovascular disease, psychiatric conditions including substance use disorders, and neurologic disease (eg, stroke). Groups defined by pain with arthritis, liver

TABLE 2. *(Continued)* Patient-Clustering Studies From High-risk Populations[27-31,33-38]

Citation; N and source population	Study population year(s)	High-risk selection criterion	Segmentation method (model performance metrics)	Included variables	Model selection method	Resulting patient groups
Multimorbidity population (*continued*)						
Bayliss et al (2019)[34]; 9617 US adult Kaiser Permanente members ≥65 years who completed at least 1 MRA	2014-2017	Identified as having "advanced illness" based on complex or multiple chronic conditions or geriatric syndromes	Cluster and LCA (goodness-of-fit metrics not reported)	20 patient-reported items from the MRA, including health status, psychosocial factors, mood symptoms, health behaviors, ADL/iADLs	Cluster analysis: cubic clustering criteria and pseudo F-statistic, then clinical judgment. LCA: Bayesian information criteria then clinical judgment.	The cluster analysis resulted in 14 clusters, including: •Healthy (22%) •Inactive but relatively healthy (8%) •Memory, hearing, balance (6%) •Balance, urinary incontinence, pain (5%) •Pain with poor sleep, but active (4%) •Poor physical and mental health (4%) •Globally sick (3%) •Memory, ADL/IADL limitations (3%) •Not easily interpreted (6 clusters, 45%) LCA resulted in an 8-class solution, 5 clinically interpretable classes (64% of the population) and 3 classes not easily clinically interpretable (36% of the population).
Guisado-Clavero et al (2018)[30]; 190,108 adults aged 65-94 years who used primary health care centers in Barcelona, Spain	2009	2 or more comorbid chronic conditions	k-means cluster analysis stratified by age and sex (Jaccard bootstrapping values)	83-85 chronic conditions with at least 1% prevalence in each subpopulation	Calinski-Harabasz criteria	6 cluster solution for each analysis: •Nonspecific (34%-38%) •MSK (14%-18%) •Endocrine-metabolic (13%-20%) •GI (women)/GI-respiratory (men) (10%-16%) •Neuropsychiatric (9%-15%) •CVD (6%-12%)
Violán et al (2018)[31]; 408,994 patients aged 45-64 years in Catalonia, Spain	2010	2 or more comorbid chronic conditions	k-means cluster analysis stratified by sex (Jaccard bootstrapping values and exclusivity)	263 diagnoses from EHR	Calinski-Harabasz index	Women: •Nonspecific (41%) •MSK (15%) •Skin, eye and ear, breast (13%) •GI, liver, kidney stones (12%) •Metabolic, HTN, and heart (11%) •Infections, injuries, GU (10%) Men: •Nonspecific (39%) •Drug abuse, liver, lung, mood (15%) •GI, GU, venous/lymph (12%) •MSK (12%) •Heart, metabolic, retinal (11%) •Eye, ENT, skin, infections (11%)

ACG, adjusted clinical group; ADL, activities of daily living; ADRD, Alzheimer disease and related disorders; BIC, Bayesian information criteria; CAN-2-H, Care Assessment Needs score; CHF, congestive heart failure; CKD, chronic kidney disease; CLD, chronic liver disease; COPD, chronic obstructive pulmonary disease; CVA, cerebrovascular accident; CVD, cardiovascular disease; DM, diabetes mellitus; EHR, electronic health record; ENT, ear, nose, and throat; ESRD, end-stage renal disease; GERD, gastroesophageal reflux disease; GI, gastrointestinal; GU, genitourinary; HBV, hepatitis B virus; HCV, hepatitis C virus; HTN, hypertension; IADL, independent ADL; IHD, ischemic heart disease; IRT, item response theory; LBP, low back pain; LCA, latent class analysis; MRA, Medicare Risk Assessment; MSK, musculoskeletal; NH, nursing home; PTSD, posttraumatic stress disorder; SDOH, social determinants of health; SNF, skilled nursing facility; SUD, substance use disorder; TIA, transient ischemic attack.

[a]Social risks included food insecurity, housing safety, financial, employment, health literacy, and social support.

[b]The 12 clinical variables included presence of specific complex chronic conditions (IHD, CHF, atrial fibrillation, ADRD, COPD/asthma, and CKD), presence of ≥2 complex chronic conditions, presence of ≥6 chronic conditions, SNF or NH use, use of home care services, hospitalization, and functional limitations.

TABLE 3. Diagnosis-Clustering Studies Among High-risk Patient Populations[24-26,32]

Citation; N and source population	Study population year(s)	High-risk selection criterion	Segmentation method (model performance metrics)	Included variables	Model selection method	Resulting diagnosis groups
Multimorbidity population						
Cornell et al (2009)[24]; 1,327,382 US adults receiving primary care in the Veterans Health Administration	1997-2000	2 or more comorbid chronic conditions	Hierarchical cluster analysis (goodness-of-fit metrics not reported)	45 chronic condition diagnoses from EHR, reduced to 23 to improve clinical interpretability of resulting clusters	Clinical judgment including clinical interpretability	• Obesity (obesity, LBP, OA, BPH, GERD) • Metabolic (DM, HTN, HLD, IHD) • Neurovascular (PVD, CVA, TIA, ADRD, seizures) • Liver (HBV, HCV, CLD, HIV) • Dual diagnosis (SUD, alcohol, schizophrenia, bipolar) • Mixed anxiety-depression (depression, PTSD, anxiety)
Deruaz-Luyet et al (2017)[26]; 888 Swiss adults using primary care	2015	3 or more comorbid chronic conditions	Hierarchal cluster analysis (goodness-of-fit metrics not reported)	24 chronic condition diagnoses from EHR	No criteria given	• Cardiovascular • Metabolic and age-related • Tobacco, alcohol, and COPD • Pain, MSK, and psychological
Roso-Llorach et al (2018)[32]; 408,994 adults aged 45-64 years in Catalonia, Spain	2010	2 or more comorbid chronic conditions	Sex-stratified hierarchal cluster (approximate unbiased probability) and EFA (goodness-of-fit metrics not reported)	79 (women) and 73 (men) active acute/ chronic diagnoses from EHR	Cluster analysis: Rand index, pseudo T-squared statistic EFA: Scree-plot with the "elbow rule"	Cluster analysis resulted in 53 clusters for women, 15 clusters for men. EFA resulted in 9 factors for women and 10 factors for men. Neither analysis was easily clinically interpretable.
Foguet-Boreu et al (2015)[25]; 322,328 adults ≥ 65 years using primary care in Catalonia, Spain	2010	2 or more comorbid chronic conditions	Sex- and age-stratified cluster analyses (approximate unbiased probability)	263 diagnoses from EHR	Adjusted Rand Index, then clinical judgment	Numbers of clusters per stratum ranged from 42 to 85. There were a small number of clinically interpretable clusters, but the majority were not easily clinically interpretable.

ADRD, Alzheimer disease and related disorders; BPH, benign prostatic hyperplasia; CLD, chronic liver disease; COPD, chronic obstructive pulmonary disease; CVA, cerebrovascular accident; DM, diabetes mellitus; EFA, exploratory factor analysis; EHR, electronic health record; GERD, gastroesophageal reflux disease; HBV, hepatitis B virus; HCV, hepatitis C virus; HLD, hyperlipidemia; HTN, hypertension; IHD, ischemic heart disease; LBP, low back pain; MSK, musculoskeletal; OA, osteoarthritis; PTSD, posttraumatic stress disorder; PVD, peripheral vascular disease; SUD, substance use disorder; TIA, transient ischemic attack.

disease, obesity, cancer, and renal disease were also found in more than 1 study. It is important to note that in groups named for their most common 1 or 2 conditions, patients typically have additional health conditions beyond those that the group is named for, and some may not have the condition the group is named for but rather other conditions in a pattern similar to that of other patients in the same group.

DISCUSSION

In this structured literature review, we found 12 articles that applied empiric segmentation methods to high-risk or high-cost patient populations. Nine articles were published since 2018, indicating rapidly growing interest in these techniques. Underlying populations were most often defined by multimorbidity and less often by high observed or predicted cost or utilization. Most studies segmented patients based on clinical conditions, but a few incorporated functional or social conditions. Groups defined by cardiometabolic, mental health, substance use disorder, and neurologic conditions were common. However, groupings varied in ways that could be traced to health care system context and selection of population and modeling inputs. This suggests that although model results from other systems may provide a starting point for common segments to be expected, systems will want to conduct segmentation analyses on their own high-risk patient populations to obtain the most relevant and applicable results. Based on these observations, we created a "prescription" for health care systems intending to use data-driven segmentation for their high-risk patient populations (**Figure**).

Our findings indicate that the results and applicability of segmentation analyses are heavily influenced by study design decisions. Setting will influence the type of data available and interventions available to use for resulting patient segments. For studies set within a health care system, all but 1 included only those patients engaged with primary care over time. This may reflect how health care systems define patients who "belong" to them, but it may also reflect that primary care providers are tasked

with managing multimorbidity and mitigating hospitalization risk within these systems.[43] High-risk selection criteria similarly affect the available data and interventions. Analyzing high-cost patients will highlight conditions that are expensive and not necessarily common, such as sepsis, whereas analyzing those with multimorbidity will highlight conditions that represent a significant proportion of disease burden among the population, regardless of whether they drive utilization.

The studies reviewed paint a compelling picture of how important the selection of inputs is in determining model results. Simply put, you cannot get out what you do not (or cannot) put in. Most studies included only clinical diagnoses, limiting findings to patterns among those conditions. Studies adding patient-reported health status or social risks found groupings distinguished by those variables. One study combined diagnostic data with utilization data, and resulting groups had a combination of those factors.[33] Therefore, it is essential in segmentation analyses to carefully preselect the input variables that will most directly inform the resulting tailored interventions.

Segmentation model results are also highly dependent on how input variables were measured. For example, if diagnoses are based on billing codes, conditions that are often left off billing claims will be underrepresented. If efficiency is attained by using easily accessed data, it needs to be balanced against improved accuracy that can be obtained through more comprehensive but cumbersome data extraction (eg, indicators that incorporate lab results, prescription records, or patient-reported information). Further, bias inherent to data sources will affect segmentation results. For example, people in the United States who belong to minority racial groups are less likely to receive a depression diagnosis from a health care provider[44]; thus, segmentation results may not reveal the importance of depression in that population. A balance needs to be struck between efficiency and completeness of input measures, and sources of bias should be considered when interpreting and applying results.

The major choice in modeling approach is among models that build patient clusters based on latent similarities among their traits (eg, latent class analysis) and models that build condition (or other trait) clusters. A thorough discussion of model types in each category, and their pros and cons, has been well summarized in other literature.[22,45,46] Essentially, patient clustering results in patients being "assigned" to groupings that they most resemble. In contrast, condition clustering will assign each condition to a cluster and most patients will cross multiple clusters. Intended application can drive the decision among approaches, such as dividing a population for assignment to case management vs developing interventions open to any patient with a certain cluster of conditions. In this review, more studies used a patient clustering approach, and these studies' results were more clinically interpretable.

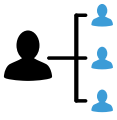
Regardless of the segmentation methods used, it is important to evaluate model fit using appropriate metrics. Statistical metrics of model performance, when carefully chosen and interpreted, can provide information about how well the model fits the underlying data, but they should supplement and not supplant utility in model selection. A model with better statistical metrics that results in a less useful population segmentation may not be preferred over an alternative model with inferior metrics but more actionable segments. Relatedly, special attention should be paid to the areas of the population with lower measures of model fit to help understand

FIGURE 1. Prescription for Designing a Data-Driven Segmentation Project to Inform Care Planning and Interventions for High-risk Populations

POPULATION: Ensure that the population is large enough to support the data-driven modeling approach. Consider whether the population is heterogeneous enough to warrant designing different interventions for subpopulations. Consider whether to define the population based on high predicted risk, high observed cost, or multimorbidity.

SITUATION: Identify the outcome of interest (eg, improved health care utilization, health-associated quality of life, quality-of-care metrics). Consider the range of potential interventions and domains over which the interventions could potentially address or could be tailored, as this will define the lens through which the segmentation will proceed.

DATA INPUTS: Assess available data sources to identify what information is available and how it maps to the information needed to develop and tailor the interventions in step 2. Consider collecting new information if needed or acknowledge where the limits of available information will constrain the segmentation models and influence the results. Acknowledge any bias inherent in the data and consider how this may influence the model results.



MODELING APPROACH: Perform the population segmentation using either a patient or health condition clustering approach. In either case, develop a method in advance for how to determine the clinically optimal solution. This will include statistical/technical approaches but should also include clinical/expert review to choose the most useful result among top contenders. Consider adjusting the inputs and rerunning the segmentation analysis to explore how choice of inputs affects the results.

NEXT STEPS: Describe resulting patient groups in terms of sociodemographics, utilization patterns, and modifiable health risks and health care needs. Validate groupings based on prospective health outcomes. Design interventions tailored to each group's profile.

where the modeling approach may be failing. The ultimate measure of success is demonstration of improved outcomes after applying interventions informed by the modeling results.

Notable limitations to this review include that (1) the time period was restricted to 20 years, although a hand search for additional studies found only 1 relevant study published prior to 2010; (2) studies with geriatric populations, general-risk populations, or disease-defined populations were not included; (3) gray literature was not included; and (4) studies published in languages other than English were not included. Formal meta-analysis of results was not possible because of the variety of study contexts and techniques that influenced results.

After using a thoughtful approach to high-risk population segmentation, key next steps are to (1) describe the characteristics of patients in each group, including demographics, utilization patterns, modifiable risk factors, and gaps in receipt of recommended care, and (2) demonstrate that resulting groups have clinically distinct outcomes. Both are integral to validating clinical significance of groups and identifying care needs and appropriate intervention settings for each group. We identified only 2 studies that prospectively examined observed utilization or outcomes occurring after patient segmentation.[33,47] Once segments are determined, risk prediction models can also be tailored to the factors that predict outcomes within each group; we did not find any studies that took this approach.

The ultimate goal of patient segmentation models should be to inform clinical care and system-level care planning. We did not find any studies that reported the results of patient- or system-level interventions based on empiric segments of high-risk patients. Future intervention studies can use patient segments as a starting point for design of tailored interventions, using descriptive and qualitative methods to further identify the health care needs, common care settings, and outcome-associated risk factors to address for each group. Trajectories of risk, utilization, or outcomes observed for each patient segment can inform study design to account for each group's expected amount of regression to the mean.[48,49] In many segmentation approaches, there are patients who remain unassigned to a group—these patients can be prioritized as likely requiring more individualized needs assessments. In addition to patient-level interventions, health care systems can also use segmentation results for system-level planning, such as staffing and training required to meet each group's needs.[10] Responsibility for care coordination or panel management can be assigned to the person in the system most appropriate for each group. Finally, learning health care systems can use their robust data and advanced analytic capabilities to track, and iteratively improve, care delivery and outcomes for each high-risk patient group.

CONCLUSIONS

We found significant variation in the data choices and modeling approaches among studies that empirically segmented high-risk or high-cost patients and their health conditions. The decisions made during the modeling process greatly affected the groups that emerged from models. Careful consideration should be given to modeling design to maximize the utility of the resulting groups in health care system program design. The next wave of studies segmenting high-risk patients can use lessons learned from this review to improve study design and take the next steps of using clinical validation and needs assessment to design tailored interventions for patient groups. Ultimately, interventions tailored to meet the needs of empirically derived segments of high-risk patient populations need to be tested to determine whether this approach will improve patient outcomes. ■

Acknowledgments

The authors thank Evelyn Chang, MD, MSHS; Denise Deverts, PhD; Stephen Fihn, MD, MPH; and Donna Zulman, MD, MS, for comments on earlier versions of this manuscript.

Author Affiliations: Division of General Internal Medicine, University of Pittsburgh (JA, AMR), Pittsburgh, PA; VA Center for Health Equity Research and Promotion, VA Pittsburgh Healthcare System (JT, JM-M, AMR), Pittsburgh, PA; Division of Pharmaceutical Outcomes and Policy, University of North Carolina Eshelman School of Pharmacy (JT), Chapel Hill, NC.

Source of Funding: This work was undertaken as part of the Department of Veterans Affairs' Primary Care Analytics Team (XVA-41-061). Funding for the Primary Care Analytics Team is provided by the Department of Veterans Affairs Office of Primary Care. The funder was not involved in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; nor decision to submit the manuscript for publication. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs, the University of Pittsburgh, or the University of North Carolina.

Author Disclosures: The authors report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (JA, AMR); acquisition of data (JA, JM-M, AMR); analysis and interpretation of data (JA, JT, JM-M, AMR); drafting of the manuscript (JA, JT, JM-M, AMR); critical revision of the manuscript for important intellectual content (JA, JT, AMR); statistical analysis (AMR); obtaining funding (AMR); administrative, technical, or logistic support (JA, JM-M, AMR); and supervision (AMR).

Address Correspondence to: Jonathan Arnold, MD, MS, MSE, Division of General Internal Medicine, University of Pittsburgh, 200 Lothrop St, Pittsburgh, PA 15213. Email: arnoldjd@pitt.edu.

REFERENCES

1. Cohen SB, Yu W. The concentration and persistence in the level of health expenditures over time: estimates for the U.S. population, 2008-2009. Agency for Healthcare Research and Quality statistical brief No. 354. January 2012. Accessed November 30, 2018. https://www.meps.ahrq.gov/data_files/publications/st354/stat354.shtml
2. Zulman DM, Pal Chee C, Wagner TH, et al. Multimorbidity and healthcare utilisation among high-cost patients in the US Veterans Affairs Health Care System. *BMJ Open*. 2015;5(4):e007771. doi:10.1136/bmjopen-2015-007771
3. Schoeman JA, Chockley N. Understanding U.S. health care spending: NIHCM Foundation data brief July 2011. Kaiser Family Foundation. July 2011. Accessed June 17, 2020. https://www.kff.org/wp-content/uploads/sites/3/2014/03/nihcm-costbrief-email.pdf
4. Niles J, Litton T, Mechanic R. An initial assessment of initiatives to improve care for high-need, high-cost individuals in accountable care organizations. *Health Affairs*. April 11, 2019. Accessed June 3, 2019. https://www.healthaffairs.org/do/10.1377/hblog20190411.143015/full/
5. Long P, Abrams M, Milstein A, et al, eds. *Effective Care for High-Need Patients: Opportunities for Improving Outcomes, Value, and Health*. National Academy of Medicine; 2017. Accessed December 3, 2018. https://nam.edu/wp-content/uploads/2017/06/Effective-Care-for-High-Need-Patients.pdf
6. Edwards ST, Peterson K, Chan B, Anderson J, Helfand M. Effectiveness of intensive primary care interventions: a systematic review. *J Gen Intern Med*. 2017;32(12):1377-1386. doi:10.1007/s11606-017-4174-z
7. Zulman DM, Pal Chee C, Ezeji-Okoye SC, et al. Effect of an intensive outpatient program to augment primary care for high-need Veterans Affairs patients: a randomized clinical trial. *JAMA Intern Med*. 2017;177(2):166-175. doi:10.1001/jamainternmed.2016.8021
8. Yoon J, Chang E, Rubenstein LV, et al. Impact of primary care intensive management on high-risk veterans' costs and utilization: a randomized quality improvement trial. *Ann Intern Med*. 2018;168(12):846-854. doi:10.7326/M17-3039

9. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)*. 2014;33(7):1123-1131. doi:10.1377/hlthaff.2014.0041
10. O'Malley AS, Rich EC, Sarwar R, et al. How accountable care organizations use population segmentation to care for high-need, high-cost patients. *Issue Brief (Commonw Fund)*. 2019;2019:1-17.
11. Leppin AL, Gionfriddo MR, Kessler M, et al. Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. *JAMA Intern Med*. 2014;174(7):1095-1107. doi:10.1001/jamainternmed.2014.1608
12. 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(2):185-208; discussion 209-212. doi:10.1111/j.1468-0009.2007.00483.x
13. Rudin RS, Gidengil CA, Predmore Z, Schneider EC, Sorace J, Hornstein R. Identifying and coordinating care for complex patients: findings from the leading edge of analytics and health information technology. RAND Corp. 2016. Accessed December 6, 2019. https://www.rand.org/pubs/research_reports/RR1234.html
14. Joynt KE, Figueroa JF, Beaulieu N, Wild RC, Orav EJ, Jha AK. Segmenting high-cost Medicare patients into potentially actionable cohorts. *Healthc (Amst)*. 2017;5(1):62-67. doi:10.1016/j.hjdsi.2016.11.002
15. Clough JD, Riley GF, Cohen M, et al. Patterns of care for clinically distinct segments of high cost Medicare beneficiaries. *Healthc (Amst)*. 2016;4(3):160-165. doi:10.1016/j.hjdsi.2015.09.005
16. Dumbreck S, Flynn A, Nairn M, et al. Drug-disease and drug-drug interactions: systematic examination of recommendations in 12 UK national clinical guidelines. *BMJ*. 2015;350:h949. doi:10.1136/bmj.h949
17. Chrischilles E, Schneider K, Wilwert J, et al. Beyond comorbidity: expanding the definition and measurement of complexity among older adults using administrative claims data. *Med Care*. 2014;52(suppl 3):S75-S84. doi:10.1097/MLR.0000000000000026
18. Breland JY, Asch SM, Slightam C, Wong A, Zulman DM. Key ingredients for implementing intensive outpatient programs within patient-centered medical homes: a literature review and qualitative analysis. *Healthc (Amst)*. 2016;4(1):22-29. doi:10.1016/j.hjdsi.2015.12.005
19. Piette JD, Kerr EA. The impact of comorbid chronic conditions on diabetes care. *Diabetes Care*. 2006;29(3):725-731. doi:10.2337/diacare.29.03.06.dc05-2078
20. Zulman DM, Martins SB, Liu Y, et al. Using a clinical knowledge base to assess comorbidity interrelatedness among patients with multiple chronic conditions. *AMIA Annu Symp Proc*. 2015;2015:1381-1389.
21. Vuik SI, Mayer EK, Darzi A. Patient segmentation analysis offers significant benefits for integrated care and support. *Health Aff (Millwood)*. 2016;35(5):769-775. doi:10.1377/hlthaff.2015.1311
22. 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(1):121. doi:10.1186/s12874-018-0584-9
23. 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(4):e0195243. doi:10.1371/journal.pone.0195243
24. Cornell JE, Pugh JA, Williams JW Jr, et al. Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database. *Appl Multivar Res*. 2009;12(3):163-182. doi:10.22329/amr.v12i3.658
25. Foguet-Boreu Q, Violán C, Rodriguez-Blanco T, et al. Multimorbidity patterns in elderly primary health care patients in a South Mediterranean European region: a cluster analysis. *PLoS One*. 2015;10(11):e0141155. doi:10.1371/journal.pone.0141155
26. Déruaz-Luyet A, N'Goran AA, Senn N, et al. Multimorbidity and patterns of chronic conditions in a primary care population in Switzerland: a cross-sectional study. *BMJ Open*. 2017;7(6):e013664. doi:10.1136/bmjopen-2016-013664
27. Davis AC, Shen E, Shah NR, et al. Segmentation of high-cost adults in an integrated healthcare system based on empirical clustering of acute and chronic conditions. *J Gen Intern Med*. 2018;33(12):2171-2179. doi:10.1007/s11606-018-4626-0
28. Prenovost KM, Fihn SD, Maciejewski ML, Nelson K, Vijan S, Rosland AM. Using item response theory with health system data to identify latent groups of patients with multiple health conditions. *PLoS One*. 2018;13(11):e0206915. doi:10.1371/journal.pone.0206915
29. Buja A, Claus M, Perin L, et al. Multimorbidity patterns in high-need, high-cost elderly patients. *PLoS One*. 2018;13(12):e0208875. doi:10.1371/journal.pone.0208875
30. Guisado-Clavero M, Roso-Llorach A, López-Jimenez T, et al. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. *BMC Geriatr*. 2018;18(1):16. doi:10.1186/s12877-018-0705-7
31. Violán C, Roso-Llorach A, Foguet-Boreu Q, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. *BMC Fam Pract*. 2018;19(1):108. doi:10.1186/s12875-018-0790-x
32. Roso-Llorach A, Violán C, Foguet-Boreu Q, et al. Comparative analysis of methods for identifying multimorbidity patterns: a study of 'real-world' data. *BMJ Open*. 2018;8(3):e018986. doi:10.1136/bmjopen-2017-018986
33. Keeney T, Belanger E, Jones RN, Joyce NR, Meyers DJ, Mor V. High-need phenotypes in Medicare beneficiaries: drivers of variation in utilization and outcomes. *J Am Geriatr Soc*. 2020;68(1):70-77. doi:10.1111/jgs.16146
34. Bayliss EA, Ellis JL, Powers JD, Gozansky W, Zeng C. Using self-reported data to segment older adult populations with complex care needs. *EGEMS (Wash DC)*. 2019;7(1):12. doi:10.5334/egems.275
35. Rogers A, Hu YR, Schickedanz A, Gottlieb L, Sharp A. Understanding high-utilizing patients based on social risk profiles: a latent class analysis within an integrated health system. *J Gen Intern Med*. 2020;35(7):2214-2216. doi:10.1007/s11606-019-05510-9
36. Wang L, Porter B, Maynard C, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. *Med Care*. 2013;51(4):368-373. doi:10.1097/MLR.0b013e31827da95a
37. Starfield B, Kinder K. Multimorbidity and its measurement. *Health Policy*. 2011;103(1):3-8. doi:10.1016/j.healthpol.2011.09.004
38. Schickedanz A, Sharp A, Hu YR, et al. Impact of social needs navigation on utilization among high utilizers in a large integrated health system: a quasi-experimental study. *J Gen Intern Med*. 2019;34(11):2382-2389. doi:10.1007/s11606-019-05123-2
39. Gelbard R, Goldman O, Spiegler I. Investigating diversity of clustering methods: an empirical comparison. *Data Knowl Eng*. 2007;63(1):155-166. doi:10.1016/j.datak.2007.01.002
40. Collins LM, Lanza ST. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Wiley; 2010.
41. Fabrigar LR, Wegener DT. *Exploratory Factor Analysis*. Oxford University Press; 2012.
42. Love BC. Comparing supervised and unsupervised category learning. *Psychon Bull Rev*. 2002;9(4):829-835. doi:10.3758/BF03196342
43. Chang ET, Zulman DM, Nelson KM, et al. Use of general primary care, specialized primary care, and other Veterans Affairs services among high-risk veterans. *JAMA Netw Open*. 2020;3(6):e208120. doi:10.1001/jamanetworkopen.2020.8120
44. Shao Z, Richie WD, Bailey RK. Racial and ethnic disparity in major depressive disorder. *J Racial Ethn Health Disparities*. 2016;3(4):692-705. doi:10.1007/s40615-015-0188-6
45. Wood RM, Murch BJ, Betteridge RC. A comparison of population segmentation methods. *Oper Res Health Care*. 2019;22:100192. doi:10.1016/j.orhc.2019.100192
46. Magidson J, Vermunt JK. Latent class models. In: Kaplan D, ed. *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Sage Publications, Inc; 2004:175-198.
47. Ibarra-Castillo C, Guisado-Clavero M, Violan-Fors C, Pons-Vigués M, López-Jiménez T, Roso-Llorach A. Survival in relation to multimorbidity patterns in older adults in primary care in Barcelona, Spain (2010–2014): a longitudinal study based on electronic health records. *J Epidemiol Community Health*. 2018;72(3):185-192. doi:10.1136/jech-2017-209984
48. Chang ET, Piegari R, Wong ES, Rosland AM, Fihn SD, Vijan S, Yoon J. Which patients are persistently high-risk for hospitalization? *Am J Manag Care*. 2019;25(9):e274-e281.
49. Wong ES, Yoon J, Piegari RI, Rosland AM, Fihn SD, Chang ET. Identifying latent subgroups of high-risk patients using risk score trajectories. *J Gen Intern Med*. 2018;33(12):2120-2126. doi:10.1007/s11606-018-4653-x