7-12-2016

Applying Sequential Analytic Methods to Self-Reported Information to Anticipate Care Needs

Elizabeth A. Bayliss
Kaiser Permanente Colorado Institute for Health Research; Department of Family Medicine, University of Colorado School of Medicine, elizabeth.bayliss@kp.org

J. David Powers
Kaiser Permanente Colorado Institute for Health Research, j.david.powers@kp.org

Jennifer L. Ellis
Kaiser Permanente Colorado Institute for Health Research, jenn.l.ellis@kp.org

Jennifer Barrow
Kaiser Permanente Colorado Institute for Health Research, jennifer.c.barrow@kp.org

See next pages for additional authors

Follow this and additional works at: http://repository.edm-forum.org/egems

Applying Sequential Analytic Methods to Self-Reported Information to Anticipate Care Needs

Abstract

Purpose: Identifying care needs for newly enrolled or newly insured individuals is important under the Affordable Care Act. Systematically collected patient-reported information can potentially identify subgroups with specific care needs prior to service use.

Methods: We conducted a retrospective cohort investigation of 6,047 individuals who completed a 10-question needs assessment upon initial enrollment in Kaiser Permanente Colorado (KPCO), a not-for-profit integrated delivery system, through the Colorado State Individual Exchange. We used responses from the Brief Health Questionnaire (BHQ), to develop a predictive model for receiving care in the top 25% for cost, then applied cluster analytic techniques to identify different high cost subpopulations. Per-member-per-month cost was measured from 6-12 months following BHQ response.

Results: BHQ responses significantly predictive of high cost care included self-reported health status, functional limitations, medication use, presence of 0-4 chronic conditions, self-reported ED use during the prior year, and lack of prior insurance. Age, gender, and deductible-based insurance product were also predictive. The largest possible range of predicted probabilities of being in the top 25% of cost was 3.5% to 96.4%. Within the top cost quartile, examples of potentially actionable clusters of patients included those with high morbidity, prior utilization, depression risk and financial constraints; high morbidity, previously uninsured individuals with few financial constraints; and relatively healthy, previously insured individuals with medication needs.

Conclusions: Applying sequential predictive modeling and cluster analytic techniques to patient-reported information can identify subgroups of individuals within heterogeneous populations who may benefit from specific interventions to optimize initial care delivery.

Authors

Elizabeth A Bayliss, *Kaiser Permanente Colorado Institute for Health Research; Department of Family Medicine, University of Colorado School of Medicine*; J. David Powers, *Kaiser Permanente Colorado Institute for Health Research*; Jennifer L Ellis, *Kaiser Permanente Colorado Institute for Health Research*; Jennifer Barrow, *Kaiser Permanente Colorado Institute for Health Research*; Mary Jo Strobel, *Kaiser Permanente Colorado, Department of Complete Health Solutions*; Arne Beck, *Kaiser Permanente Colorado Institute for Health Research*.

eGEMs
Generating Evidence & Methods
to improve patient outcomes

Applying Sequential Analytic Methods to Self-Reported Information to Anticipate Care Needs

Elizabeth A. Bayliss, MD, MSPH;[i,ii] J. David Powers, MS;[i] Jennifer L. Ellis, MSPH, MBA;[i] Jennifer C. Barrow, MA;[i] MaryJo Strobel, MSN, MBA;[iii] Arne Beck, PhD[i]

ABSTRACT

Purpose: Identifying care needs for newly enrolled or newly insured individuals is important under the Affordable Care Act. Systematically collected patient-reported information can potentially identify subgroups with specific care needs prior to service use.

Methods: We conducted a retrospective cohort investigation of 6,047 individuals who completed a 10-question needs assessment upon initial enrollment in Kaiser Permanente Colorado (KPCO), a not-for-profit integrated delivery system, through the Colorado State Individual Exchange. We used responses from the Brief Health Questionnaire (BHQ), to develop a predictive model for cost for receiving care in the top 25 percent, then applied cluster analytic techniques to identify different high-cost subpopulations. Per-member, per-month cost was measured from 6 to 12 months following BHQ response.

Results: BHQ responses significantly predictive of high-cost care included self-reported health status, functional limitations, medication use, presence of 0–4 chronic conditions, self-reported emergency department (ED) use during the prior year, and lack of prior insurance. Age, gender, and deductible-based insurance product were also predictive.The largest possible range of predicted probabilities of being in the top 25 percent of cost was 3.5 percent to 96.4 percent. Within the top cost quartile, examples of potentially actionable clusters of patients included those with high morbidity, prior utilization, depression risk and financial constraints; those with high morbidity, previously uninsured individuals with few financial constraints; and relatively healthy, previously insured individuals with medication needs.

Conclusions: Applying sequential predictive modeling and cluster analytic techniques to patient-reported information can identify subgroups of individuals within heterogeneous populations who may benefit from specific interventions to optimize initial care delivery.

[i]Kaiser Permanente Colorado Institute for Health Research, [ii]Department of Family Medicine, University of Colorado School of Medicine, [iii]Kaiser Permanente Colorado, Department of Complete Health Solutions

Introduction

The Patient Protection and Affordable Care Act (ACA) resulted in 11.6 million individuals purchasing insurance coverage through federal or state exchange marketplace plans and over 5 million individuals enrolling in expanded Medicaid and CHIP programs by mid-2015.[1,2] Because the marketplace is not static, and benefit and subsidy eligibility may fluctuate, many people continue to move between insurers even after initial enrollment. As a result, insurers, clinicians, and health care delivery systems continue to engage new members and patients.

To optimize health outcomes in this environment, it is increasingly important to anticipate care needs of new patients. However, predicting care needs of new beneficiaries before they utilize services is difficult. Usual approaches to assessing service needs based on past utilization and morbidity burden are not possible for individuals who have not yet received care. It is also difficult to extrapolate from published estimates of the morbidity burden of newly insured individuals because national estimates apply only to previously uninsured populations, not the larger number of transitioning beneficiaries.[3,4]

The most straightforward approach to understanding the care needs of new patients is to ask them. Prediction rules to identify high-need patients from self-reported information have been previously developed from national survey data and Medicaid needs assessments.[5,6] However, applying such predictive models in clinical settings is limited by generalizability beyond the specific setting and by population heterogeneity: Existing prediction rules may not accurately identify high risk individuals when applied in new populations, and those identified are likely to have a wide range of needs. To be actionable, effective interventions to optimize initial care delivery must be based on specific clinical and psychosocial information, and must focus on smaller target populations.

To address these limitations, we combined two analytic methods with a 10-question needs assessment, the Brief Health Questionnaire (BHQ), to identify actionable target populations for care management of newly enrolled members of an integrated delivery system. We developed predictive models to estimate the individual probability of high service needs, and combined this with cluster analytic techniques to identify meaningful subgroups within the high-need population that may be amenable to specific interventions or actions. Our premise was that enhancing traditional predictive modeling with data-based algorithms and basing both on current self-reported information would provide more accurate information on the specific care needs of new members. Specifically, smaller and more homogeneous target populations would potentially be amenable to tailored care management interventions.

Our aims in this paper were the following: (1) to identify BHQ responses and other factors among a cohort of 2014 new non-Medicare members enrolling through the Colorado State individual exchange marketplace that predict higher costs, and (2) to determine whether there are potentially meaningful clusters of BHQ responses that are associated with higher costs.

Methods

Study Setting and Population

This retrospective cohort study was conducted at Kaiser Permanente Colorado (KPCO), a not-for-profit integrated health care delivery system. Our focus for these analyses was on members with new KPCO health care benefits effective between January 1 and June 30, 2014 with no prior membership, who enrolled as individuals through Connect for Health Colorado, the Colorado state ACA marketplace ("the exchange"), and who completed the Brief Health Questionnaire in the first six months of 2014. We

excluded anyone who had a face-to-face encounter in the KPCO system prior to completing the BHQ. The index date was the date of BHQ completion. If individuals changed insurance products during the year, we selected the product around the BHQ date. KPCO offers a variety of insurance products: traditional health maintenance organization (HMO), deductible coinsurance (DHMO), and high-deductible health (HDHP) plans with health savings accounts.

The KPCO Institutional Review Board reviewed the protocol for the BHQ program and associated analyses and determined that it met criteria for an operations intervention with intent to publish rather than human subjects research. Thus it was exempted from IRB review.

Data Sources and Measures

The BHQ consists of 10 previously validated questions that were used in analyses, the responses to which, for purposes of clustering and predictive modeling, were coded 0 or 1, with 1 indicating potentially greater care needs.[7-18] The questions, with positive responses indicated in parentheses, are the following: general health status (poor/fair health); conditions interfere with daily activity (yes); asthma (yes); diabetes (yes); heart disease (yes); high blood pressure (yes); prescription medications (yes); positive Patient Health Questionnaire (PHQ)-2 depression screen (score 3 or more); financial constraints (yes); prior year inpatient and ED use (1 or more times); and no prior year insurance coverage (no insurance for more than eight months). Operating under the assumption that positive responses to the BHQ would potentially trigger clinical action, we grouped missing responses with nonpositive responses. The BHQ was offered by telephone to *eligible members* calling for an appointment and was accessible on the KPCO patient portal site. There was also limited outreach to new members who had not yet contacted the

system. Responses to BHQ questions were entered into the member's electronic health record (EHR).

We obtained utilization-related costs for members from the time of BHQ completion through the end of 2014 or disenrollment. Costs came from KPCO's costs accounting system, which combines medical utilization with the General Ledger and calculates direct unit costs and relative overhead to provide a total cost for a medical service. We divided members' total costs by their months of enrollment to create a per-member, per-month (PMPM) cost.

Analysis

The dependent variable for predictive model development was high cost of care, defined as being in the top 25 percent *of the study population for total cost of care during the time between* BHQ completion and either the end of enrollment or December 31, 2014—whichever came first. The characteristics considered for independent variables were BHQ responses, age, gender, and benefit plan. Because one of our objectives was to identify the strength and precision of BHQ responses and other readily available demographic and insurance characteristics in predicting inclusion in the top quartile of costs, we used predictive methods, as opposed to explanatory methods.[19]

We split the data into approximately equal halves to create temporal training and validation data sets. We then estimated a logistic regression model on the training data set. We evaluated all of the BHQ responses for inclusion in the regression model, as well as age, gender, and product type. We then plotted receiver operating curves (ROC) and calculated c-statistics on the training and combined data sets, estimated using the final selected model. We also calculated the c statistic on the validation data set scored using the coefficients from the model estimated on the training sample in order to get an estimate of optimism.[19]

Calibration plots under the final model were created using the same data set and estimation setup as above, where scatter plots of average actual probability and 95 percent confidence interval bands by model prediction decile were generated. These plots allow for easy visual inspection of model accuracy overall and by different ranges of model predictions.

The sample size was fixed because of the retrospective design. Assuming a requirement of 10–20 outcome events relative to candidate predictors, using a random sample of 3,034 members to develop the predictive model yielded 830 events, which allowed us to consider 41 to 83 degrees of freedom.[20]

Clustering Methods

We used k-means clustering, a partitive clustering method, to identify subgroups of patients within the highest quartile of cost that could potentially benefit from tailored outreach efforts. This method attempts to maximize the between-cluster sum of squares or minimize the within-cluster sum of squares; we used the latter option to optimize within-cluster similarity based on input variables. We used the cubic clustering criterion (CCC) and pseudo-F (PSF) statistics, as well as clinical judgment, to inform the final number of clusters. Inputs to the algorithm were BHQ questions plus type of benefit plan.

We initially developed the clusters using a cohort of members in the top 25 percent of cost; inputs were BHQ responses. We then applied the cluster algorithm to the entire study cohort and compared the distribution of the clusters to cost quintiles. This provided a picture of how much the cluster definitions were driven by cost. This comparison also helped to ensure that potential discrepancies in cluster distribution between the actual high-cost members and members who were predicted to have high cost did not eliminate any clusters or make them less useful.

Results

Six thousand forty-seven members met the criteria for this study; 1,512 were in the top 25 percent of costs. Six percent of the population (n=382) incurred no 2014 costs subsequent to their BHQ. Enrollment in an exchange plan subsequent to the BHQ ranged from 1 (0.2 percent) to 12 (33.8 percent) months. Characteristics of the population can be found in Table 1. Those incurring higher costs were more likely to be female and older, to have HMO coverage, and to have positive responses to BHQ questions. Thirteen percent of responses related to PHQ2 and prior insurance were missing, 1 percent were missing responses to the financial considerations question; other questions had 0.2 percent to 0.5 percent missing responses.

The final model results in the combined training and validation data sets with adjusted odds of predicting high cost of care are illustrated in Figure 1 and detailed in Table 2. Variables most strongly predictive of high cost of care included having fair or poor health, requiring prescription medications, and enrolling in a non-high deductible plan. Prior ED utilization, a history of being uninsured, and number (out of four) of comorbid conditions, age, and female gender were also highly significant. The c-statistic for the resulting model estimated in the combined training and validation data sets was 0.7466. After correcting for optimism, the final estimate of future model performance was 0.7080. Figure A1 in the supplemental online Appendix shows the receiver operating characteristic curves for the final model estimated in the training sample and scored in the training and validation samples to depict how discrimination may be attenuated across all predicted values in applying the model to future samples. Figure A2 in the online Appendix illustrates how the final model predictive accuracy may be maintained across most of the predictive range in applying the model to future

samples, and the potential for overprediction in the top two deciles.

In determining the optimum number of clusters in which to partition members of the top 25th percentile of costs, the CCC values suggested 2, 4, or 8 clusters, while the PSF values suggested 2 or 4. We felt that more segmentation might yield more clinically meaningful groups and selected the 8 cluster solution; characteristics of these clusters are shown in Table 3. Each cluster displays a somewhat unique combination of characteristics that might indicate a need for intervention and outreach or, conversely, no immediate needs other than preventive care.

In comparing the distribution to cost quintiles, it became apparent that three of the clusters (2, 3, and 8) were more heavily composed of high-cost members. The distribution of clusters in the top quintile of predicted probability for high cost compared with the distribution in the actual top 25 percent cost members is fairly similar, with the exception of cluster 7 being underrepresented in the top quintile of model scores (Figure A3). *Sensitivity analyses in which missing data were grouped with positive rather than nonpositive responses did not change the results of the predictive models or the clusters*.

Discussion

The newly insured Individual Exchange population encompasses subpopulations with a broad range of care needs.[7,21] Our operational and analytic initiative illustrates how complementary analytic methods coupled with the systematic collection of patient-reported data can inform care delivery for heterogeneous and potentially complex patient populations—minimizing adverse effects of discontinuity and improving system efficiencies.

Traditional approaches to anticipating care needs for complex patients have relied on predictive models to identify candidates for interventions.[14,22,23] Although valuable for estimating the probability of a given outcome for individuals, using predictive models alone to anticipate population care needs has two primary limitations:Models may identify large numbers of at-risk individuals that are too great for effective interventions; and outcomes of interest to systems and policymakers (e.g., cost of care, hospital admission) reflect many different clinical scenarios that require different intervention strategies.[24] Large-scale care management interventions such as Medicare demonstration projects are costly, labor intensive, and have varying results.[25-27] Thus, more focused approaches to care management have been recommended.[28] Tailoring successful interventions to optimize efficient and effective care requires either focused predictive models designed around specific interventions (e.g., dialysis prevention in individuals with renal disease) or additional means of identifying meaningful subpopulations.[24,29]

When applied to populations at risk for specific outcomes, cluster algorithms can identify actionable subpopulations to inform intervention development and application. Using historical data, we developed a predictive model and associated decision tool to identify individuals at risk for high cost of care. We then applied cluster techniques to partition the highest-risk group into discrete subgroups groups with identifiable care needs and in doing so created an informal typology of subpopulations. The decision tool and typology can then be used to subset new patients from similar populations into actionable subgroups. Figure 1 illustrates how historical data can be used to develop both predictive models and population typologies that can then be applied to individual level data to tailor care.

Previous applications of predictive modeling to patient-reported data have identified self-reported utilization, prescription drug use, health-related quality of life (HRQOL), and information on medical

Table 1. Characteristics of New Exchange Members Responding to the Brief Health Questionnaire Between January 1, 2014 and June 30, 2014 (N=6,047) by Low Versus High Costs

MEMBER CHARACTERISTIC	LOWEST 3 QUARTILES[1] (N=4535), N (%)	HIGHEST QUARTILE[1] (N=1512), N (%)	P VALUE
Female gender	2,574 (56.8)	979 (64.7)	<0.0001
Age group, years			
0–18	723 (15.9)	60 (4.0)	
19–34	1,148 (25.3)	298 (19.7)	
35–54	1,458 (32.2)	532 (35.2)	
55+	1,206 (26.6)	622 (41.1)	<0.0001
Product type			
HMO	1,082 (23.9)	603 (39.9)	
Deductible coinsurance	1,434 (31.6)	535 (35.4)	
High deductible health plan with HSA	2,019 (44.5)	374 (24.7)	<0.0001
Months of exchange enrollment post-BHQ			
1–3	56 (1.2)	27 (1.8)	
4–6	80 (1.8)	41 (2.7)	
7–9	1,870 (41.2)	706 (46.7)	
10–12	2,529 (55.8)	738 (48.8)	<0.0001
Poor or fair health	192 (4.2)	246 (16.3)	<0.0001
Condition that interferes with daily activity	542 (12.0)	431 (28.5)	<0.0001
Current prescription medication	1,600 (35.3)	954 (63.1)	<0.0001
One or more chronic condition	875 (19.3)	596 (39.4)	<0.0001
Number of chronic conditions (0–4)			
0	3,660 (80.7)	916 (60.6)	
1	772 (17.0)	463 (30.6)	
2	96 (2.1)	117 (7.7)	
3	7 (0.2)	11 (0.7)	
4	0 (0.0)	5 (0.3)	<0.0001
Asthma	263 (5.8)	157 (10.4)	<0.0001
Diabetes	139 (3.1)	168 (11.1)	<0.0001
Heart disease	62 (1.4)	80 (5.3)	<0.0001
High blood pressure	521 (11.5)	345 (22.8)	<0.0001
Prior year ED use	443 (9.8)	290 (19.2)	<0.0001
Prior year inpatient admission	148 (3.3)	114 (7.5)	<0.0001
Reported financial constraint	266 (5.9)	229 (15.1)	<0.0001
PHQ2 score of 3 or more	166 (3.7)	127 (8.4)	<0.0001
No insurance >8 months in prior year	1,314 (29.0)	606 (40.1)	<0.0001

Note: Members were required to have no face-to-face encounters prior to BHQ completion.

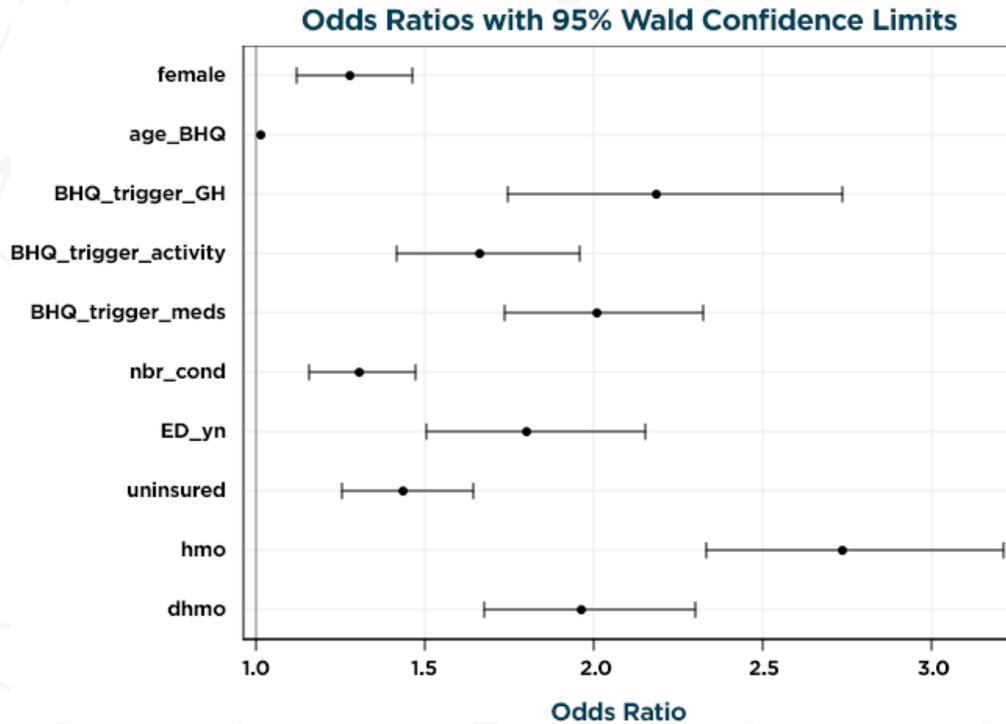**Figure 1. Final Model Results on Combined Training and Validation Data Sets**

Odds Ratios with 95% Wald Confidence Limits

Table 2. Predictive Model: Odds of Being in the Top 25% of Cost Based on Demographics and BHQ Responses*

PARAMETER	ODDS RATIO	95% CONFIDENCE	INTERVAL	P-VALUE
Female vs. male gender	1.28	1.121	1.462	0.0003
Age at BHQ administration (per year)	1.017	1.013	1.021	<.0001
Fair/poor health (vs. excellent/very good/ good)	2.185	1.747	2.734	<.0001
Condition interferes with daily activity	1.664	1.415	1.956	<.0001
Any current prescription medication	2.007	1.734	2.324	<.0001
Number of comorbid conditions (per condition)	1.306	1.158	1.472	<.0001
ED utilization in prior year	1.801	1.506	2.154	<.0001
Previously Uninsured	1.435	1.253	1.642	<.0001
HMO (HDHP referent)	2.736	2.332	3.211	<.0001
DHMO (HDHP referent)	1.963	1.676	2.298	<.0001

Notes: *C-statistic from full data model: 0.7466 (0.7324, 0.7608). See also Supplemental Digital Content, Figure A1.

conditions as predictive of cost and hospital utilization.[5,6] Our predictive model additionally identified self-reported health status, functional limitations, positive depression screen, lack of prior insurance, age, gender, and having HMO and deductible-based insurance as potential predictors of high-cost care. This information was moderately predictive of being in the top quartile for cost of care with a c-statistic of 0.75.

When this population was further partitioned with cluster techniques, these variables illustrated several potentially actionable subgroups of new members. For example, Cluster 3 (Table 3) identified 96 individuals, of whom a majority had fair/poor health, required prescription medications, had financial constraints, were likely to have a positive depression screen, and had high hospital utilization during the prior year. Members of this group are likely to benefit from prompt and relatively intensive care management outreach with special attention to their behavioral health needs. In contrast, Cluster 6 identified 219 individuals who were in good health, but all of whom had no insurance during the preceding year—potential users of "catch-up" preventive and other care. Members of the largest group, Cluster 7, are notable for relatively low morbidity and low prior utilization, but moderate medication needs. Targeting care management resources to this large group would be less efficient, but specific pharmacy services might be useful and welcome. *Individuals identified in Cluster 5 who report relatively low morbidity and high emergency service use may benefit from convenient access to urgent care services*.

Because cluster methods are inherently exploratory, not all clusters necessarily provide actionable information. In our example, cluster 1 comprises individuals who appear to have been previously insured, have at least one chronic condition, and require prescription medications. They may be transitioning between insurers, shopping primarily on price through the exchange, or have undetected medical needs. As some of these assumptions require additional verification, this cluster may not be ideal for initial intervention development.Close collaboration with *local* operational partners is essential to interpret results and to apply the method iteratively with a mind to effective interventions. Ideally, input variables are selected to reflect potential interventions—such as identifying those with financial needs when community or other social support resources are available. In our application, inputs were limited to items from a needs assessment, but the sequential process of identifying a larger at-risk group using a predictive model and then partitioning the group into actionable subgroups can be applied with any available and relevant variables.

Previous studies have combined partitioning techniques and multivariable models to identify subgroups with different clinical prognoses or alone to identify subgroups with different clinical risks and needs.[30-32] Decision trees have also been used to select variables for predictive models.[22] We are unaware of any applications of predictive modeling, cluster techniques, and patient-reported data, or of applying these techniques to inform care management. Although our approach may appear burdensome, current analytic and data collection capacities facilitate combining techniques. When indicated, patient-reported data are increasingly available and extractable.

This initiative has several limitations. First, our sample of BHQ responses primarily reflected individuals who contacted KPCO for an appointment. Thus the variables that predict high cost of care and the associated typology of new members reflect users of the delivery system and may not incorporate subpopulations of nonusers. Beneficiaries with high deductible health plans were less likely to be in the

Figure 2. Applying Sequential Analytic Methods to Inform Tailored Care Delivery
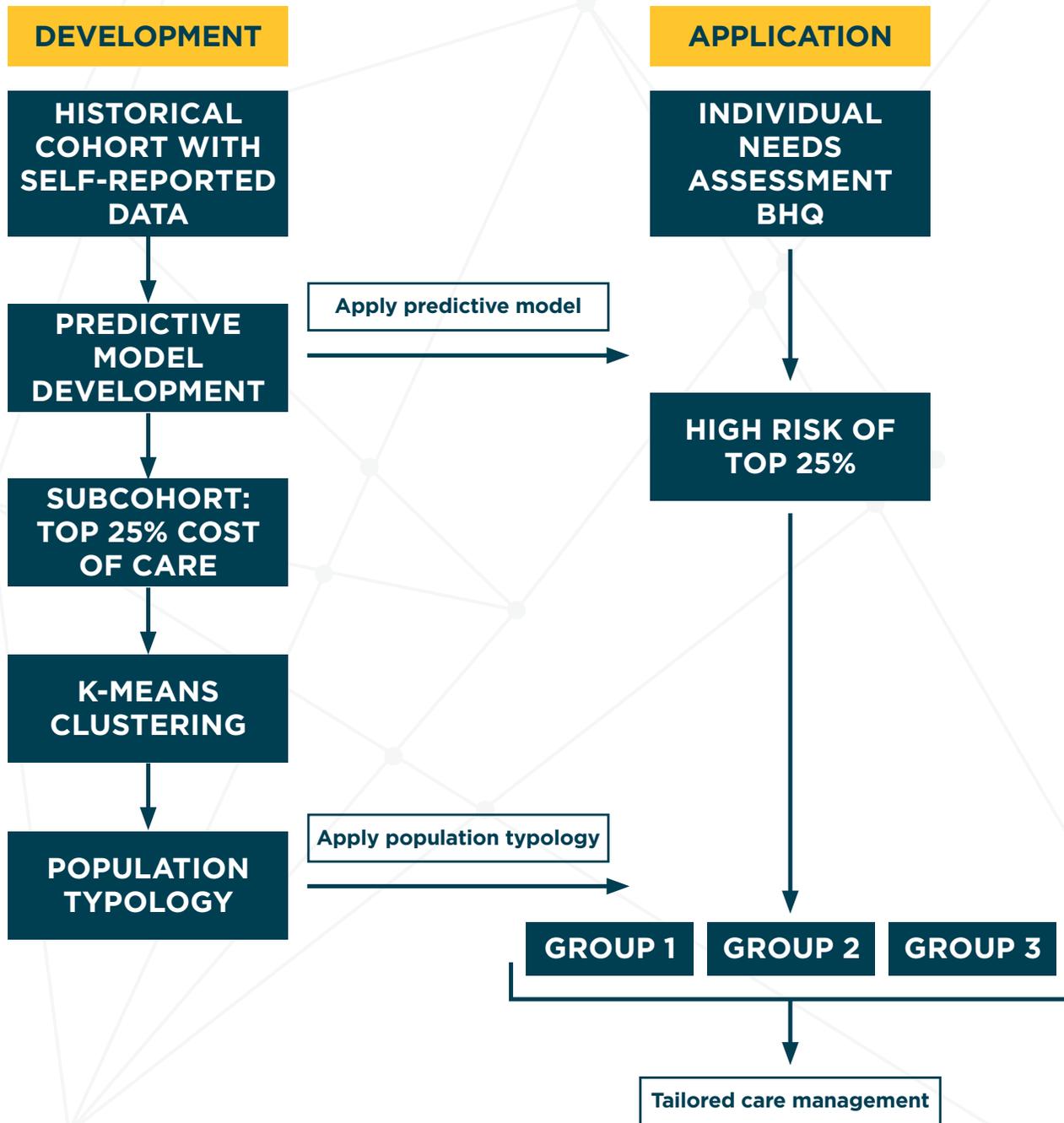
Table 3. Clusters Within the Top 25th Percentile of Costs (% with Characteristic) (N = 1,512)

	CLUSTER							
Characteristic	1 N=244	2 N=141	3 N=96	4 N=89	5 N=128	6 N=219	7 N=499	8 N=96
Fair/poor health	0%*	24%**	74%**	100%**	11%*	0%*	1%*	32%**
Condition interferes with daily activity	23%*	100%**	92%**	22%*	31%	12%*	0%*	61%**
Current prescription medication	89%**	58%*	96%**	79%**	52%*	88%**	36%*	56%*
1+ chronic condition	100%**	0%*	91%**	78%**	0%*	74%**	0%*	35%**
Reported financial constraint	9%*	10%*	59%**	9%*	4%*	9%*	2%*	100%**
3+ on PHQ2	3%*	12%**	44%**	10%	6%*	4%*	2%*	28%**
Prior year ED use	17%	0%*	71%**	9%*	100%**	12%*	0%*	19%
Prior year admission	6%	4%*	51%**	1%*	22%**	3%*	2%*	0%*
No insurance >8 months in prior year	0%*	39%	19%*	55%**	34%*	100%**	25%*	99%**
DESCRIPTORS, NOT INCLUDED IN CLUSTERING ALGORITHM								
Age group								
0–18	0%*	6%**	0%*	0%*	6%**	0%*	8%**	0%*
19–34	14%*	21%	11%*	7%*	31%**	11%*	26%**	22%
35–54	27%*	35%	40%**	43%**	42%**	35%	36%	30%*
55+	58%**	38%	47%**	51%**	20%*	54%**	29%*	48%**
Female gender	59%*	67%	66%	63%	70%**	62%	68%	60%*
Plan type								
DHMO	36%	30%*	35%	30%*	42%**	31%*	37%	40%**
HDHP	24%	21%*	17%*	21%*	25%	23%	30%**	18%*
HMO	40%	49%**	48%**	48%**	33%*	46%**	33%*	43%

Note: *Below population average 99% CL. **Above population average 99% CL.

top cost quartile and may be underrepresented. This may reflect self-selection of lower cost plans by healthier individuals or deferred care due to financial constraints (or both). Developing insurance products that optimize access and cost while encouraging individual self-efficacy is necessary for clarifying population health care needs and utilization patterns.Second, our assessment was limited to new members of a single integrated delivery system. Using similar methodologies across other delivery systems and settings (or using statewide public health survey data) could provide a broader picture of population care needs and inform more population-level interventions.Third, it could be argued that using a predictive model as a first step is not necessary, as data mining methods can theoretically identify any number of discrete clusters, given a set of input variables.[33] However, we felt that applying a decision tool to identify individuals at risk and subsequently segment them based on a derived typology was more consistent with a traditional approach to care delivery—for managing the health of individuals within populations, the two methods are complementary. Finally, this report focuses on the analytic aspects of the BHQ initiative. However, identifying high need individuals is the easy part. Developing and implementing successful (and tailored) interventions requires clinical and operational engagement, an appreciation of system and patient context, and effective allocation of resources.

Conclusion

Delivering optimal care to complex patient populations requires a range of analytic methods to understand and address population heterogeneity, as well as patient-reported information to capture health and functional status, financial needs, and other nuanced information. Our example illustrates one approach of applying predictive modeling and cluster algorithms to patient-reported data.

Population health leaders might apply these methods with patient-reported or other data to identify actionable subgroups for early care management, quantify high need populations for resource allocation, tailor insurance products to optimize access, or develop self-efficacy interventions for patients with specific care needs. Anticipating care needs before they happen can improve system efficiency and minimize the adverse health outcomes associated with discontinuity.

Acknowledgements

References

1. US Department of Health and Human Services. Health Insurance Marketplaces 2015 Open Enrollment Period: March Enrollment Report. ASPE Issue Brief. 2015:1-72.
2. Artiga S, Rudowitz R, Gates A, Snyder L. Recent trends in Medicaid and CHIP enrollment as of January 2015: early findings from the CMS perfomrance indicatgor project. Washington, DC: 2015.
3. Decker SL, Kostova D, Kenney GM, Long SK. Health status, risk factors, and medical conditions among persons enrolled in Medicaid vs uninsured low-income adults potentially eligible for Medicaid under the Affordable Care Act. JAMA. 2013;309(24):2579-86.
4. Forrest CB, Lemke KW, Bodycombe DP, Weiner JP. Medication, diagnostic, and cost information as predictors of high-risk patients in need of care management. The American Journal of Managed Care. 2009;15(1):41-8.
5. Wherry LR, Burns ME, Leininger LJ. Using Self-Reported Health Measures to Predict High-Need Cases among Medicaid-Eligible Adults. Health Services Research. 2014;49(S2):2147-72.
6. Leininger LJ, Friedsam D, Voskuil K, DeLeire T. Predicting high-need cases among new Medicaid enrollees. The American Journal of Managed Care. 2014;20(9):e399-407.

7. Bayliss EA, Ellis JL, Strobel MJ, McQuillan DB, Petsche IB, Barrow JC, et al. Characteristics of Newly Enrolled Members of an Integrated Delivery System after the Affordable Care Act. The Permanente Journal. 2015;19(3):4.

8. DeSalvo KB, Fan VS, McDonell MB, Fihn SD. Predicting mortality and healthcare utilization with a single question. Health Services Research. 2005;40(4):1234-46.

9. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Medical Care. 2003;41(11):1284-92.

10. Dorr DA, Jones SS, Burns L, Donnelly SM, Brunker CP, Wilcox A, et al. Use of Health-Related, Quality-of-Life Metrics to Predict Mortality and Hospitalizations in Community-Dwelling Seniors. Journal of the American Geriatrics Society. 2006;54(4):667-73.

11. de Boer AG, Wijker W, de Haes HC. Predictors of health care utilization in the chronically ill: a review of the literature. Health Policy. 1997;42(2):101-15.

12. Donnan PT, Dorward DW, Mutch B, Morris AD. Development and validation of a model for predicting emergency admissions over the next year (PEONY): a UK historical cohort study. Archives of Internal Medicine. 2008;168(13):1416-22.

13. Halfon N, Newacheck PW, Wood DL, St Peter RF. Routine emergency department use for sick care by children in the United States. Pediatrics. 1996;98(1):28-34.

14. Wallace E, Stuart E, Vaughan N, Bennett K, Fahey T, Smith SM. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Medical Care. 2014;52(8):751.

15. Ginde AA, Lowe RA, Wiler JL. Health insurance status change and emergency department use among US adults. Archives of Internal Medicine. 2012;172(8):642-7.

16. Li AK, Covinsky KE, Sands LP, Fortinsky RH, Counsell SR, Landefeld CS. Reports of financial disability predict functional decline and death in older patients discharged from the hospital. Journal of General Internal Medicine. 2005;20(2):168-74.

17. Bindman AB, Chattopadhyay A, Auerback GM. Interruptions in Medicaid coverage and risk for hospitalization for ambulatory care–sensitive conditions. Annals of Internal Medicine. 2008;149(12):854-60.

18. Baker DW, Sudano JJ, Albert JM, Borawski EA, Dor A. Lack of health insurance and decline in overall health in late middle age. New England Journal of Medicine. 2001;345(15):1106-12.

19. Harrell Jr F. Jr. Regression modeling strategies. New York: Springer; 2001.

20. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.

21. Donohue JM, Papademetriou E, Henderson RR, Frazee SG, Eibner C, Mulcahy AW, et al. Early Marketplace Enrollees Were Older And Used More Medication Than Later Enrollees; Marketplaces Pooled Risk. Health Affairs. 2015:10.1377/hlthaff.2015.0016.

22. Shadmi E, Flaks-Manov N, Hoshen M, Goldman O, Bitterman H, Balicer RD. Predicting 30-day readmissions with preadmission electronic health record data. Medical Care. 2015;53(3):283-9.

23. Hong CS, Atlas SJ, Ashburner JM, Chang Y, He W, Ferris TG, et al. Evaluating a model to predict primary care physician-defined complexity in a large academic primary care practice-based research network. Journal of General Internal Medicine. 2015;30(12):1741-7.

24. Wallace E, Smith SM, Fahey T, Roland M. Reducing emergency admissions through community based interventions. BMJ. 2016;352:h6817.

25. Fout B, Weinberg D, Bill N, Kahn K, Sennett C. Evaluation of the Senior Risk Reduction Demonstration (SRRD) Under Medicare. Columbia, MD: Centers for Medicare and Medicaid Services, 2013 December 20, 2013.

26. Peikes D, Chen A, Schore J, Brown R. Effects of care coordination on hospitalization, quality of care, and health care expenditures among Medicare beneficiaries: 15 randomized trials. JAMA. 2009;301(6):603-18.

27. McCarthy D, Ryan J, Klein S. Models of care for high-need, high-cost patients: An evidence synthesis. 2015 October, 2015.

28. Anderson GF, Ballreich J, Bleich S, Boyd C, DuGoff E, Leff B, et al. Attributes common to programs that successfully treat high-need, high-cost individuals. The American Journal of Managed Care. 2014;21(11):e597-e600.

29. Weiss JW, Platt RW, Thorp ML, Yang X, Smith DH, Petrik A, et al. Predicting Mortality in Older Adults with Kidney Disease: A Pragmatic Prediction Model. Journal of the American Geriatrics Society. 2015;63(3):508-15.

30. Crocetti E, Mangone L, Scocco GL, Carli P. Prognostic variables and prognostic groups for malignant melanoma. The information from Cox and Classification And Regression Trees analysis: an Italian population-based study. Melanoma Research. 2006;16(5):429-33.

31. Hougham GW, Ham SA, Ruhnke GW, Schulwolf E, Auerbach AD, Schnipper JL, et al. Sequence Patterns in the Resolution of Clinical Instabilities in Community-Acquired Pneumonia and Association with Outcomes. Journal of General Internal Medicine. 2014;29(4):563-71.

32. Marschollek M, Gövercin M, Rust S, Gietzelt M, Schulze M, Wolf K-H, et al. Mining geriatric assessment data for in-patient fall prediction models and high-risk subgroups. BMC Medical Informatics and Decision Making. 2012;12(1):1.

33. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics. 2008;77(2):81-97.

Appendix: Supplemental Digital Content

Figure A1. Receiver Operating Characteristic Curves for Final Model Estimated In Training Sample and Scored in Training and Validation Samples
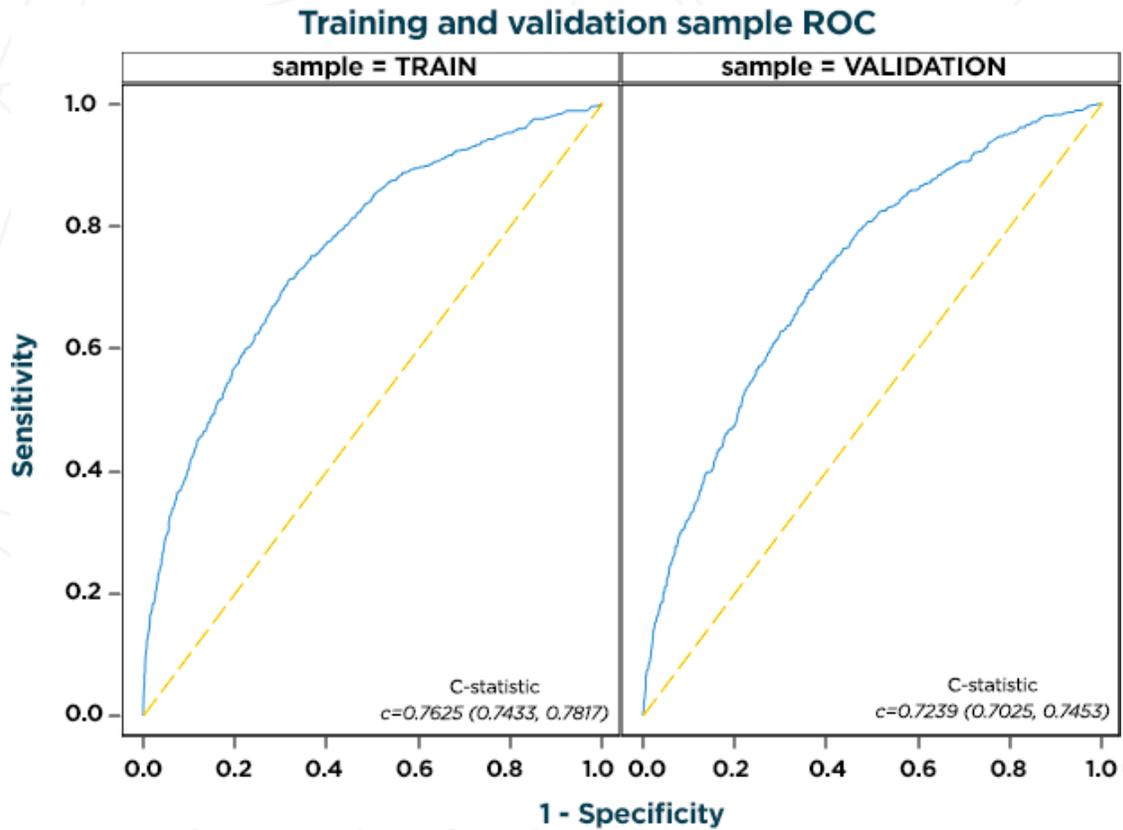
LINE	DESCRIPTION	C-STATISTIC (95% CI)
1	Model built and scored on training data set	0.7625 (0.7433,0.7817)
2	Model built on training data set, scored on validation	0.7239 (0.7025,0.7453)
3	Optimism (Line 1–Line 2)	0.0386 (0.0408, 0.0364)
4	Model built and scored on combined data set	0.7466 (0.7324, 0.7608)
5	Final corrected (Line 4–Line 3)	0.7080 (0.6916, 0.7244)

Figure A2. Calibration of Final Model Estimated in Training Sample and Scored in Training and Validation Samples
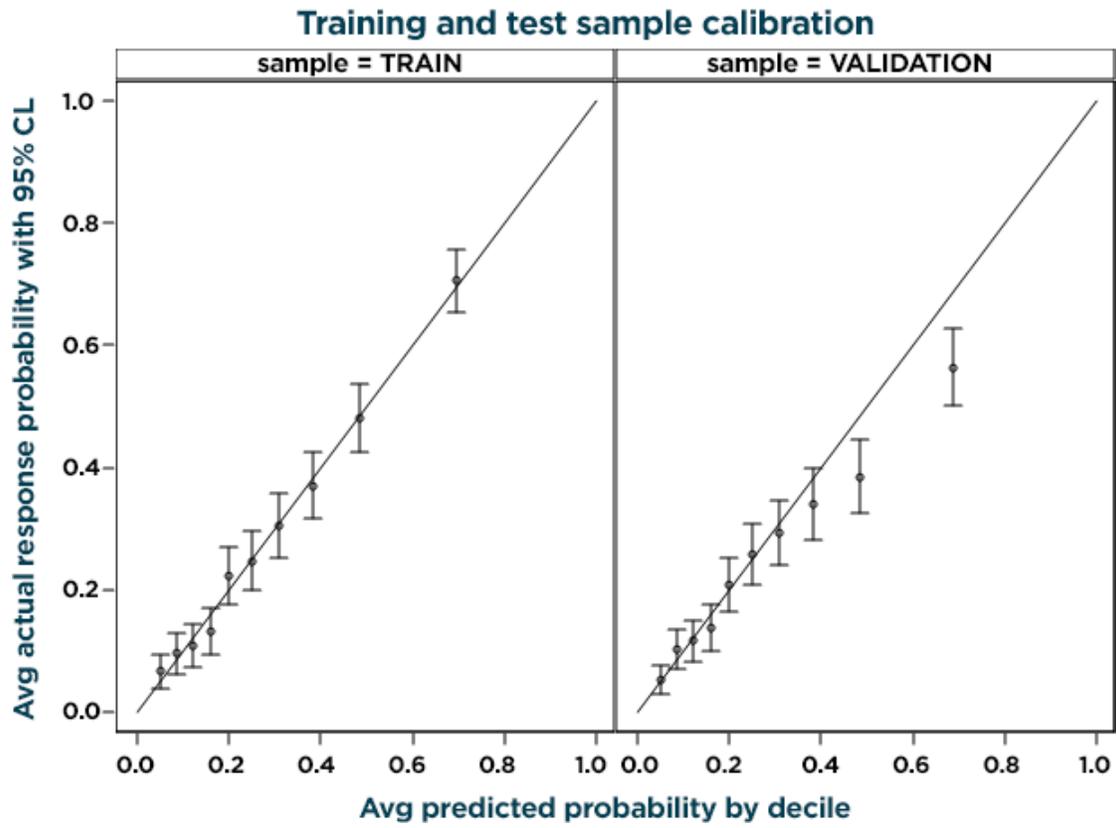

Training and test sample calibration

Figure A3. Comparison of Cluster Distribution Between Highest Model Quintile to Actual Top 25% Cost Population

Compare cluster distribution in highest model quintile to actual top 25% cost