Sample selection in the face of design constraints: Use of clustering to define sample strata for qualitative research

Lane F. Burgette PhD¹ | José J. Escarce MD, PhD² | Susan M. Paddock PhD³ |
Marjorie S. Ridgely JD³ | Warren G. Wilder MPhil⁴ | Dolores Yanagihara MPH⁵ |
Cheryl L. Damberg PhD³

¹RAND Corporation, Pittsburgh, Pennsylvania

²University of California at Los Angeles, Los Angeles, California

³RAND Corporation, Santa Monica, California

⁴Balyasny Europe Asset Management LLP, London, UK

⁵Integrated Healthcare Association, Oakland, California

Correspondence

Lane F. Burgette, PhD, RAND Corporation, Pittsburgh, Pennsylvania.
Email: burgette@rand.org

Funding information

National Institutes of Health, Grant/Award Number: R33AG044252

Objective: To sample 40 physician organizations stratified on the basis of longitudinal cost of care measures for qualitative interviews in order to describe the range of care delivery structures and processes that are being deployed to influence the total costs of caring for patients.

Data Sources: Three years of physician organization-level total cost of care data (n = 156 in California) from the Integrated Healthcare Association's value-based pay-for-performance program.

Study Design: We fit total cost of care data using mixture and K-means clustering algorithms to segment the population of physician organizations into sampling strata based on 3-year cost trajectories (ie, cost curves).

Principal Findings: A mixture of multivariate normal distributions can classify physician organization cost curves into clusters defined by total cost level, shape, and within-cluster variation. K-means clustering does not accommodate differing levels of within-cluster variation and resulted in more clusters being allocated to unstable cost curves. A mixture of regressions approach focuses overly on anomalous trajectories and is sensitive to model coding.

Conclusions: Statistical clustering can be used to form sampling strata when longitudinal measures are of primary interest. Many clustering algorithms are available; the choice of the clustering algorithm can strongly impact the resulting strata because various algorithms focus on different aspects of the observed data.

KEYWORDS

biostatistical methods, health care costs, sampling

1 | INTRODUCTION

Large year-over-year growth in health spending, which has outpaced the growth in the general economy,^{1,2} has led policy makers to institute provider payment reforms^{3,4} in an attempt to slow the rate of cost growth (ie, bend the cost curve).⁵ Public and private payers are implementing alternative payment models—including bundled payments, value-based payments, and Accountable Care

Organizations—to change the way care is delivered and foster provider accountability for overall costs and quality of care.

The Integrated Healthcare Association's (IHA) value-based pay-for-performance (VBP4P) program is one such effort. The VBP4P program uses performance-based financial incentives, public recognition, and public reporting of performance to incentivize physician organization (PO) improvement on a common set of measures (patient experience, clinical quality, resource use,

total cost of care, and use of health IT⁶). The VBP4P program, one of the country's largest value-based purchasing experiments, targets approximately 170 POs in California representing more than 8 million covered lives in HMO and POS insurance products. IHA began implementing the VBP4P program in 2013, rewarding POs for quality and cost performance. POs can earn shared savings provided they meet thresholds for cost trend and quality performance.

Many value-based payment efforts focus providers' attention on bending the cost curve. However, it is unknown whether such programs are effective at improving quality and reducing costs, and if so, what actions providers take to improve quality and reduce costs. Our motivating study is a longitudinal evaluation of the impact of the IHA VBP4P program to assess changes in quality and cost trends, and to understand what actions POs are taking to succeed under VBP4P, particularly as it relates to controlling spending. We planned to conduct qualitative interviews with a subset of the POs exposed to the VBP4P program to understand the range of actions POs are taking to reduce costs and improve quality, identify care delivery structures and processes associated with different PO cost curves, and understand the challenges they faced in bending the cost curve. Given resource constraints, the question was how to sample a small number of POs from the universe of 156 POs that were eligible for VBP4P incentives in 2014 (POs with complete quality and TCC data for 2013 measurement year).

Our goal is to develop a sampling design to facilitate understanding of determinants of differences in total cost of care (TCC) curves across POs, by ensuring that the sampling process (a) uses data to identify important cost curve trajectory types (based on level and shape) and (b) samples sufficient numbers of POs from each trajectory type. In this context, we are interested in both the level (PO-level mean TCC) and changes in TCC (curves) across time. Because POs' responses to the value-based incentives are likely to depend on the current and recent historical TCC, stratifying the sample on the basis of types of cost curves guarantees coverage of the characteristics that the study team hypothesizes are important to understand factors associated with current and future heterogeneity in TCC values.

While our application focuses on drawing a sample of POs based on their total cost of care trajectories, these methods have broad applicability in the field of health services research regarding a general class of problems that require sampling that stratifies on longitudinal measures. Such methods could be used to inform sampling strategies in studies related to evaluating longitudinal changes in quality of care, health outcomes, and access to care. The units in our sampling frame are POs, but the methods could just as well be applied to sample among patients, hospitals, or states. We focus on several relatively easy-to-implement approaches to stratify the sampling frame into homogenous clusters, comparing their relative strengths and weaknesses. In addition to explaining the details of our study design, we provide broad guidance on issues that are likely to be important when these methods are applied in the context of other health services studies.

2 | METHODS

2.1 | Clustering methods

Overall, we are interested in clustering POs in our sample frame according to the trajectories of TCC with the ultimate goal of stratifying the sample according to TCC trajectory shape and level. The literature contains dozens of clustering methods, including some, such as developmental trajectory mixture modeling⁷ that have been applied to examining hospital-level cost and care outcomes.⁸ While few, if any, of these methods were developed for sampling purposes any of them could be used to segment the sampling frame for purposes of stratification for sample selection. However, none accommodate an important feature in our data—namely the precision of PO cost estimates varies across POs, given that PO cost estimates reflect averages of patient-level costs. PO cost estimates based on more observations will be more precisely estimated than those based on fewer observations. While some of the more advanced clustering approaches reviewed by Hastie et al⁹ could be modified to accommodate this feature, many are rather complex and computationally demanding to estimate.¹⁰ Instead—in the novel context of designing a study to examine health care cost trajectories—we focus on relatively simple clustering models that are more easily implemented and nonetheless perform well in our study, in the sense of resulting in qualitatively different groups that do not overly stratify the sample.

2.2 | Total cost of care data

TCC represents all spending for health maintenance organization (HMO) and point of service (POS) patients within a PO, including hospital, ambulatory, pharmacy, and ancillary services. TCC is estimated by computing the total cost per patient incurred by each plan and then aggregating across contracted plans to generate the average TCC per PO. The TCC is adjusted for different patient mixes using the concurrent DxCG Relative Risk Score¹¹ and for geographic differences in wages using the Centers for Medicare & Medicaid Services hospital wage index. Our study team only has access to the PO-level yearly TCC measures, not patient-level data.

2.3 | Notation

We use the following notation throughout this paper. Let $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,T})$ be the observed values that define the trajectory of unit i in our sampling frame for times t between 1 and T . In our study, $\mathbf{x}_i$ is a three-component vector of TCC observations for PO i , as measured in the three baseline years of the study. Our goal is to find cluster membership indicators m_i such that, for POs labeled i and j , $m_i = m_j$ implies that POs i and j are in the same cluster. Some of our models refer to a "center" of the cluster, which can be considered a "typical" TCC trajectory for the cluster. We denote the cluster centers as $\mathbf{c}_k$ where k indexes cluster; these have the same dimension as each of the $\mathbf{x}_i$ vectors. N denotes the number of distinct POs.

2.4 | Approach 1: K-means

K-means is perhaps the most commonly used method for applied clustering exercises. Using K-means to define sample strata dates back at least to the work of Jarque.¹² In this algorithm, the data are segmented into K distinct groups where each group has an associated cluster center, $\mathbf{c}_k$ for $k = 1, \dots, K$. The cluster centers and cluster membership indicators m_i are chosen such that the total within-cluster variance, $\sum_{i=1}^N d(\mathbf{x}_i, \mathbf{c}_{m_i})^2$, is minimized. In this expression, $d(\mathbf{x}, \mathbf{c})$ is used to denote the Euclidian distance between vectors $\mathbf{x}$ and $\mathbf{c}$, that is, the square root of the sum of squared component-wise differences, or $(d(\mathbf{x}, \mathbf{c}))^2 = (x_1 - c_1)^2 + \dots + (x_T - c_T)^2$.

Finding exact solutions to the K-means clustering problem (ie, directly minimizing $\sum_{i=1}^N d(\mathbf{x}_i, \mathbf{c}_{m_i})^2$, over $\mathbf{c}_k$ and $\{m_i\}$) is computationally demanding and scales poorly in n and K . However, approximate solutions can be quickly estimated in an iterative fashion and are widely available. In this application, we use visual assessments of clustered trajectories to choose the number of clusters K . (See also Appendix S1) Although some more automated approaches suggest higher numbers of clusters,^{13,14} the following goals lead us to choose $K = 6$. We wish to avoid using too many clusters or methods that produce clusters with few observations, as these will overstratify the sample. Also, we favor methods that produce clusters that are qualitatively different in terms of their TCC trajectories. We found that $K = 6$ balanced these goals.

2.5 | Approach 2: Mixture of multivariate normal distributions

Second, we consider a finite mixture of multivariate normal distributions. This model defines a data-generating process, such that the trajectories are produced in several steps. First, if there are K clusters, the probability that PO i is in cluster k is given by p_k for $k = 1, \dots, K$. More briefly, $\Pr(m_i = k) = p_k$ for $k = 1, \dots, K$. Then, a cluster assignment for PO i is drawn (independently of all other PO cluster assignments). Next, each cluster is assumed to have an associated multivariate normal distribution, where the means $\mathbf{c}_k$ and variance structures Σ_k of these distributions are allowed to vary between clusters. Finally, trajectory $\mathbf{x}_i$ is drawn from the distribution that corresponds to cluster m_i . One can consider models where the number of clusters grows as the sample size increases (ie, infinite or nonparametric mixture models),^{15,16} but such methods will not be considered here because such models can result in sparsely populated clusters and can be substantially more complex to implement.

The clustering model itself seeks to reverse the data-generating process. Having observed the trajectories $\mathbf{x}_i$, we wish to infer from which cluster (ie, which normal distribution) each observation arose. Along the way, we estimate the parameters of the normal distributions, though they are not of direct interest.

We note that the POs differ substantially in terms of their numbers of enrolled individuals across which to measure TCC and construct a PO average TCC, which can impact the precision of the estimate of a PO's TCC. Smaller POs may experience year-to-year

fluctuations in TCC beyond the changes that are driven by structural factors since the TCC measure is an average, and averaging over small sample sizes results in less stable means than for large sample sizes, all else being equal.

Our finite mixture model can then be written as

$$\mathbf{x}_i \sim \text{normal}(\mathbf{c}_{m_i}, \Sigma_{m_i} + \text{diag}(\sigma^2/r_{i,t}))$$

where cluster membership indicators m_i are drawn from a categorical distribution with K categories and with parameters $\mathbf{p} = (p_1, \dots, p_K)'$, that is, $\Pr(m_i = k) = p_k$. The vectors $\mathbf{c}_k$ are the centers of the clusters, describing the typical trajectory shape for each cluster. The matrices Σ_k are diagonal with period-specific variances that account for within-cluster variation about the typical trajectory. We denote the number of enrollees in PO i at time t as $r_{i,t}$. The diagonal matrix $\text{diag}(\sigma^2/r_{i,t})$ accounts for differences in stability of the TCC measures based on the number of enrollees in each PO. Including this component gives somewhat less weight to the smaller POs when defining the typical TCC trajectory shapes and, therefore, in estimating the cluster membership identifiers (m_i) themselves. Said another way, Σ describes the variance around the cluster center for an arbitrarily large PO; σ^2 can be interpreted as twice the additional within-cluster variance for a PO with r enrollees versus one with $2r$ enrollees. The vector of parameters $\mathbf{p}$ allows some clusters to be more heavily populated than others.

As with K-means, the choice of the number of clusters K is important for this model. Because a data-generating process defines the method, the selection of K can be framed as a model-selection problem. However, our ultimate goal is not to describe or predict future TCC observations, which are the typical motivations for standard model selection techniques. As with K-means, six clusters yielded clusters that were qualitatively distinct, while avoiding clusters with too few members. We take a Bayesian approach to estimating these models using the freely available Stan software package called from R.¹⁷ See Appendix S2 for code.

We note that successful clustering methods often omit features that would be important for predictive modeling of the same data. For example, we could relax the assumption that the Σ_k are diagonal to improve model fit. However, doing so encourages the clustering to be defined more by shape than level and—in this case—results in overlapping clusters that are not as qualitatively different as those that result from the model described above.

Finally, we note that K-means can be framed as a simplified version of our mixture of normal approach. Using $\mathbf{x}_i \sim \text{normal}(\mathbf{c}_{m_i}, \text{diag}(\sigma^2))$, equal multinomial probabilities (ie, $\mathbf{p} = (1/K, \dots, 1/K)$) are equivalent to the K-means approach above (eg, Hastie, Tibshirani, and Friedman, Section 14.3.7⁹). Note that this is less flexible than our mixture of normal model in three regards: (a) The multinomial probabilities are assumed to be equal; (b) the variance at each time point is assumed to be the same, rather than using an estimated diagonal variance matrix; and (c) our mixture of normal approach accommodates increased variance for smaller POs.

2.6 | Approach 3: Mixture of regression models

A motivating goal of the clustering approach was to find clusters that are defined on the basis of both level (ie, mean TCC values across the

available years) and change over time, such as increasing or decreasing cost curve trajectories. To focus explicitly on these two characteristics of the curves, we considered a regression model of the form

$$y_{i,t} = \alpha_{m_{i,1}} + t\beta_{m_{i,t}} + \varepsilon_{i,t}.$$

Here, we assume that the TCC measure for the i th PO at time t (for $t = 0, 1, 2$) is described by an intercept α (modeling the level) and a slope parameter β to model changes in TCC over time. In order to induce clustering, we assume that both the slopes and intercepts of the regression may vary by PO, but that there are only a small number of distinct slopes and intercepts. For example, if there are four distinct intercepts and two distinct slopes, the model naturally segments the sampling frame into $4 \times 2 = 8$ clusters defined by the slope/intercept combination.

In theory, this model is appealing since interpretation would seem to be straightforward. The $m_{i,1}$ indicator denotes assignment of PO i to a level-based class that corresponds to one of the differing levels of TCC at time point 1. Similarly, $m_{i,t}$ defines a clustering on rates of change. Hence, all unique combinations of slopes and intercepts would define the final clustering. For example, in the 3×2 setting, the final clusters might reasonably be interpreted as all combinations of low, moderate, and high TCC, crossed with low and high levels of growth. However, as we will see, the resulting clusters can be less easy to interpret in practice.

2.7 | Finding a “best” clustering for a given model

K-means, as typically implemented, is a “hard” clustering approach in the sense that each unit is unambiguously assigned to a single cluster: It belongs to the cluster whose center is nearest to it. The mixture of normal and mixture of regression model approaches discussed above are considered “soft” in the sense that there is a *probability* that an observation arose from a particular cluster, which we call the cluster membership probabilities; these probabilities must add to one for each observation. For example, for an observation that is “between” two of the cluster centers for the mixture of normals model, evidence will not clearly determine to which cluster it should be assigned. This will be reflected in membership probabilities that are not close to 1 for any single cluster. If clusters are well-separated from one another, observations that land near one of the cluster centers will be inferred to belong to such a cluster with probabilities near 1.

For our study application, we need a single, “best” clustering in order to define the sampling strata. For our “soft” clustering models, the raw output consists of random draws of plausible cluster membership indicators. These draws reflect probabilities that two POs are in the same cluster when averaged across a large number of such draws. For example, if a clustering model indicates that observations x_i and x_j are in the same cluster with high probability, nearly all of the generated sets of cluster indicators will have $m_i = m_j$. On the other hand, if POs labeled i and j' have 25 percent probability of being in the same cluster according to the model, only around 25 percent of the draws of cluster indicators should have $m_i = m_{j'}$.

A related issue in soft clustering models is that of “label switching.” The clustering model assumes that observations arise from clusters labeled, for example, 1, 2, 3, and 4. However, the exact same model fit can be achieved by permuting the cluster labels throughout the model from one draw to the next. For example, consider clustering a scalar x . If Cluster 1 is centered at $c_1 = -3.5$ and Cluster 2 is centered at $c_2 = 4.1$, the measures of the clustering model’s fit will be wholly unchanged by setting $c'_1 = 4.1$ and $c'_2 = -3.5$ while simultaneously switching the values of m_i from 1 to 2 and vice versa. If label switching occurs, quantities such as the probability that a PO belongs to Cluster 1 become almost meaningless. Rather, we want to find the single clustering that reflects the probabilities that each pair of observations is in the same cluster. For this, we follow the least squares approach described by Dahl, Day, and Tsai.¹⁸ First, we calculate the pairwise probabilities that each PO belongs to the same cluster. Then, among all sampled cluster configurations, we select the one that is closest to the average, with closeness being judged from the perspective of least squares. (The labeling of K-means clusters is also arbitrary but there is only a single set of labels, rather than the multiple draws of the soft clustering models).

3 | RESULTS

3.1 | K-means

Figure 1 shows the K-means results. Roughly, four of the clusters are defined on the basis of level, one cluster is defined by rapid increases in costs, and one is defined by rapid decreases. There are several issues with K-means that compromise its utility for defining sampling strata in our study. Most importantly, the variation in each cluster is assumed to be the same. Hence, clusters are only defined on the basis of their centers, c_k , and not the variance within clusters. It is also not possible to account for enrollment numbers a standard K-means framework, meaning that year-to-year noise may be more influential in the clustering. Even so, given the ubiquitous software to estimate K-means clusters and their easy interpretation, we would argue that this is a sensible starting place, even if only as an exploratory analysis.

3.2 | Mixture of normals

Figure 2 displays results of our second clustering approach, based on the mixture of multivariate normal distributions. The resulting clusters can be characterized qualitatively, with 1 through 5 correspond to increasing levels of TCC, with Cluster 6 containing a collection of anomalous TCC curves that are characterized by very large year-to-year shifts in TCC measures.

Compared to Figure 1, the treatment of the anomalous TCC curves demonstrates the flexibility of the mixture of normals model relative to standard K-means, since the mixture of normals allows differing within-cluster spreads. Whereas K-means results in two clusters that are characterized by rapid shifts over time, the mixture of normal groups all of the rapidly changing POs into one cluster, and provides greater resolution between the levels of the stable cost curves.

FIGURE 1 Physician organization cost clusters based on K-means clustering

3.3 | Mixture of regressions

Finally, we consider the mixture of regression models, which attempts to segment the data so that a simple linear regression model describes the cost curves within each cluster. While the goal of the mixture of regression models is to explicitly segment the sample by level and slope, in practice this method falls short in our application. In our primary application of this method, we allow for four distinct intercepts and two distinct slopes, yielding eight distinct clusters. We code the years as $t = 0, 1, 2$. The result of this clustering approach is shown in Figure 3. Here, there was a tendency for the slope to discriminate between the typical rate of growth (Clusters 1 through 4 in Figure 3) and rather anomalous rates of growth (Clusters 5 and 6). Two of the clusters are empty in this approach and are not included in the figure.

The second drawback is that the clustering results are very sensitive to how time is centered (or not) in the model in a manner that can degrade interpretability of the results. For instance, we also consider recording time as $t = 1, 2, 3$ in a model that is otherwise the same as the one used to produce Figure 3. In Figure 4, we focus on

two of the eight clusters for this slightly modified model. These clusters that share an estimated intercept were defined more strongly by levels than rates of change over time. Hence, with a small number of distinct intercepts allowed, the parameters that putatively describe changes over time can in essence be forced to provide more resolution on the overall levels. Fitted values in regression models are typically insensitive to recentering covariates, but this clustering focuses on estimated regression parameters. Because of this, the resulting clusters may be highly sensitive to such elements of the specification and may adversely impact the interpretability of the model. An approach that would circumvent these centering issues would be to allow a new intercept/slope combination to be drawn for each cluster. However, this is merely a more rigid version of our preferred mixture of normal, assuming linear trends over time.

3.4 | Choosing among the methods

Often, one can choose among competing clustering approaches as a simple matter of model selection. For example, one can use BIC or another model comparison criterion to choose between alternative

FIGURE 2 Physician organization cost clusters arising from the finite mixture of normals formulation

Notes: Clusters 1 through 5 are primarily defined by their level, whereas Cluster 6 is characterized by rapid changes over time. The mean within-year standard deviations of TCC for Cluster 1 through 6 are as follows: 206.8, 183.5, 213.4, 240.4, 222.6, and 764.2, respectively.

clustering models. In our case, however, we are less concerned with model fit and more concerned with the sample characteristics that result from using the clusters as sample strata.

Table 1 summarizes key differences in the clustering approaches. One of our goals was to explicitly cluster on differential rates of growth over time. The mixture of regressions is the only approach that does this. However, this results in four clusters that together contain only six POs (with two being empty). For use in a sampling algorithm, having so many sparsely populated clusters is undesirable. Moreover, potential users of the mixture of regressions approach should be aware of its sensitivity of the coding of the time dimension, which is not an issue for *K*-means or the mixture of normals.

While the *K*-means and mixture of normal properties are similar, there are key differences. First, the mixture of normals allocates one cluster to the unstable trajectories versus two for *K*-means since our implementation of the mixture of normals allows for different variances across clusters. We feel that grouping the unstable TCC trajectories in a single cluster better reflects the

goals of our sample stratification. In part, this is because unstable TCC trajectories may reflect data quality issues. The POs in Cluster 6 of Figure 2 are not so small that their unstable trajectories can be explained by simple year-to-year variation and changes in patient risk profiles do not explain the changes either. Thus, distinguishing between the two types of unstable trajectories in Figure 1 is of limited use. Additionally, we account for the different levels of enrollment across POs with the mixture of normals, which standard *K*-means cannot accommodate. Thus, a smaller PO will be less influential in defining a cluster's center for the mixture of normals. This is appropriate because we anticipate higher year-to-year variance for small POs due to random fluctuation alone. (Fitting the mixture of normals without the enrollment term results in similar results, except one low-TCC cluster is defined by low variance and another is defined by high variance, rather than having five clusters defined almost entirely by level, which we find to be preferable for our study goals.) For these reasons, we choose the mixture of normal as our preferred clustering approach.

FIGURE 3 Results of the mixture of regressions approach, using four distinct intercepts and two distinct slopes
 Note: Two empty clusters are not displayed.

All of our methods were designed to accommodate both differences in level and in shape. In the end, however, the strong majority of POs were allocated to clusters that were differentiated by level rather than slope or shape. Said another way, most of the variance in the observations is explained by starting levels rather than changes over time.

3.5 | Drawing the sample

Before drawing the sample, we consider the allocation across clusters. If interest were primarily in making cross-cluster comparisons, we might allocate the sample equally within clusters. If estimating population-level summaries were more important, we might allocate proportional to size (ie, the number of POs within each cluster). Initial study resources allowed for us to sample 30 POs; in that sample, we made a compromise between these two approaches, so that POs in small clusters (ie, those with fewer members) have a higher probability of being sampled than those in large clusters, but the large clusters have more elements sampled. Because there was concern

about the data quality of the POs in Cluster 6 of Figure 2, the study team made the decision to forgo sampling that cluster.

After initial contacts, we found that POs with fewer than 5000 commercial enrollees were not sufficiently involved with the P4P program to answer many of the questions in the interview. Hence, the sample was re-drawn to focus on POs with commercial enrollment of at least 5000. Further, because the competitiveness of various health care markets may have substantial impact on a PO's ability to enact cost changes, we rejected samples until each region (Bay Area/Sacramento, Central Valley, Inland Empire, Los Angeles, and Orange County/San Diego) had at least four sampled POs.

After sampling the 30 POs, further resources were made available to survey an additional 10 POs. These were spread across clusters to achieve a proportional-to-size allocation for the POs with at least 5000 enrollees. No further geographic constraints were required of this sample, and the POs of Cluster 6 that display unstable cost trajectories remain unsampled. When sampled POs were not willing or able to participate in the study, we drew replacements

FIGURE 4 Mixture of regressions results for two clusters with the same estimated intercept but different slopes with an alternative coding of time

	K-means	Mixture of normals	Mixture of regressions
Number of clusters allocated to unstable trajectories	2/6	1/6	4/8
Sensitive to coding of time	No	No	Yes
Linearity assumption	No	No	Yes
Constant variance assumption	Yes	No	Yes
Explicit clustering of different rates of change over time	No	No	Yes
Account for number of enrollees	No	Yes	No
Sizes of small clusters	5, 7	7, 8	0, 0, 2, 4

TABLE 1 Comparison of clustering approaches

from the pool of unsampled units in the same cluster and, if possible, geographic region.

By way of comparison, if a sample of size 30 (as was possible with our initial budget) were drawn without any stratification, there would be a 17.3 percent chance that none of the eight POs from the lowest cost cluster in the mixture of normal clustering would have been sampled, using hypergeometric distribution calculations. Even with our larger sample of size 40, there would still be an 8.8 percent chance of sampling none of the lowest-TCC POs would be sampled. Given the focus on understanding costs, the unstratified sample would be taking a rather large chance of missing an important group altogether.

4 | DISCUSSION AND CONCLUSION

Many health policy questions can only be answered by sampling a limited number of high-level professionals such as physicians or

hospital administrators. Recruiting such samples can be laborious and expensive, heightening demands to maximize the utility of the sample, given constraints on study resources. For moderate-to-small sample sizes, stratification is important for assuring that the sample reflects important population characteristics.¹⁹ When focusing on longitudinal measures—such as TCC or quality measures—it can be important to stratify the sample to guarantee coverage of important trajectory types. We are not aware of other health policy studies that have stratified the sample on the basis of clustering of longitudinal measures.

Sometimes, trajectories are defined by large numbers of observations. Such situations are common in clinical contexts²⁰ but also occur in the context of health services study, such as a study on emergency room access by Luo et al.²¹ The methods discussed here would be insufficient to model data where subjects have hundreds or even millions of observations per observational unit in the sampling frame. However, the higher-level principle remains: We could

estimate a model that segments the sampling frame into small number of groups, such that the sampling can be performed within each group.

Significant investments are being made to incentivize value in health care, so it is important to understand changes in costs and quality over time and how different providers respond to incentives. For example, Medicare is currently performing experiments to better understand and optimize use of Accountable Care Organizations (ACOs), which encourage provision of high-value care by sharing savings with providers. Claims-based studies of ACOs are feasible for large numbers of patients.^{22,23} However, in-depth interviews with ACO stakeholders may be important to optimize ACO policy. Stratifying a sample on the basis of trajectories of cost, quality, or value measures could ensure that even a modest-sized study sample would contain adequate numbers of stakeholders with different experiences. Understanding why there are differences between the best and worst performers may be able to help all achieve better results. We believe that the methods described in this paper could facilitate better study designs for in-depth studies that are aimed at understanding and improving cost, quality, and value over time.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported through grant funding from the National Institutes of Health Common Fund (5R33AG044252-04). The content and opinions expressed in this publication are solely the responsibility of the authors and do not reflect the official position of NIH or the U.S. Department of Health and Human Services. The authors report no conflicts of interest.

ORCID

Lane F. Burgette <https://orcid.org/0000-0001-5945-1760>

REFERENCES

1. Keehan SP, Cuckler GA, Sisko AM, et al. National health expenditure projections, 2014–24: spending growth faster than recent trends. *Health Aff.* 2015;34(8):1407-1417.
2. Moses H, Matheson DH, Dorsey ER, George BP, Sadoff D, Yoshimura S. The anatomy of health care in the United States. *J Am Med Assoc.* 2013;310(18):1947-1964.
3. Song Z, Rose S, Safran DG, Landon BE, Day MP, Chernew ME. Changes in health care spending and quality 4 years into global payment. *N Engl J Med.* 2014;371(18):1704-1714.
4. Obama B. United States health care reform: progress to date and next steps. *J Am Med Assoc.* 2016;316(5):525-532.
5. Cutler D. How health care reform must bend the cost curve. *Health Aff.* 2010;29(6):1131-1135.
6. Integrated Healthcare Association. Value based P4P. 2017. <http://www.ihc.org/our-work/accountability/value-based-p4p>. Accessed June 4, 2017.
7. Jones BL, Nagin DS, Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. *Sociol Methods Res.* 2001;29(3):374-393.
8. Xu X, Li SX, Lin H, et al. "Phenotyping" hospital value of care for patients with heart failure. *Health Serv Res.* 2014;49(6):2000-2016.
9. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*, vol. 2, no. 1. New York, NY: Springer; 2009.
10. Lee J, Müller P, Zhu Y, Ji Y. A nonparametric Bayesian model for local clustering with Application to Proteomics. *J Am Stat Assoc.* 2013;108(503):775-788.
11. Integrated Healthcare Association. California Pay for Performance Program Measurement Year 2013 P4P Manual. 2016. <http://www.ihc.org/resources/my-2013-vbp4p-manual>. Accessed April 12, 2016.
12. Jarque CM. A solution to the problem of optimum stratification in multivariate sampling. *Appl Stat.* 1981;30(2):163-169.
13. Fraley C, Raftery AE, Murphy TB, Scrucca L. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597, Department of Statistics, University of Washington; 2012.
14. Fraley C, Raftery AE. Model-based clustering, discriminant analysis and density estimation. *J Am Stat Assoc.* 2002;97:611-631.
15. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *J Am Stat Assoc.* 1995;90(430):577-588.
16. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. *J Am Stat Assoc.* 2006;101(476):1566-1581.
17. Stan Development Team. RStan: the R interface to Stan. R package version 2.14.1. 2016. <http://mc-stan.org/>
18. Dahl DB, Day R, Tsai JW. Distance-based probability distribution on set partitions with applications to protein structure prediction. Technical report, Texas A&M University; 2008.
19. Adams JL, Wickstrom SL, Burgess MJ, Lee PP, Escarce JJ. Sampling patients within physician practices and health plans: multistage cluster samples in health services research. *Health Serv Res.* 2003;38(6p1):1625-1640.
20. Di C-Z, Crainiceanu CM, Caffo BS, Punjabi NM. Multilevel functional principal component analysis. *Ann Appl Stat.* 2009;3(1):458-488.
21. Luo W, Cao J, Gallagher M, Wiles J. Estimating the intensity of ward admission and its effect on emergency department access block. *Stat Med.* 2013;32(15):2681-2694.
22. McWilliams JM, Chernew ME, Landon BE, Schwartz AL. Performance differences in year 1 of pioneer accountable care organizations. *N Engl J Med.* 2015;372(20):1927-1936.
23. Nyweide DJ, Lee W, Cuedon TT, et al. Association of Pioneer accountable care organizations vs traditional Medicare fee for service with spending, utilization, and patient experience. *J Am Med Assoc.* 2015;313(21):2152-2161.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Burgette LF, Escarce JJ, Paddock SM, et al. Sample selection in the face of design constraints: Use of clustering to define sample strata for qualitative research. *Health Serv Res.* 2019;54:509–517.
<https://doi.org/10.1111/1475-6773.13100>