Sample size considerations in observational health care quality studies

Sharon-Lise T. Normand\textsuperscript{1,2,*,†} and Kelly H. Zou\textsuperscript{1,3}

\textsuperscript{1}Department of Health Care Policy, Harvard Medical School, Boston, MA 02115, U.S.A.
\textsuperscript{2}Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.
\textsuperscript{3}Department of Radiology, Brigham and Women’s Hospital, Boston, MA 02115, U.S.A.

SUMMARY

A common objective in health care quality studies involves measuring and comparing the quality of care delivered to cohorts of patients by different health care providers. The data used for inference involve observations on units grouped within clusters, such as patients treated within hospitals. Unlike cluster randomization trials where often clusters are randomized to interventions to learn about individuals, the target of inference in health quality studies is the cluster. Furthermore, randomization is often not performed and the resulting biases may invalidate standard tests. In this paper, we discuss approaches to sample size determination in the design of observational health quality studies when the outcome is binary. Methods for calculating sample size using marginal models are briefly reviewed, but the focus is on hierarchical binomial models. Sample size in unbalanced clusters and stratified designs are characterized. We draw upon the experiences that have arisen from a study funded by the Agency for Healthcare Research and Quality involving assessment of quality of care for patients with cardiovascular disease. If researchers are interested in comparing clusters, hierarchical models are preferred. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: health care quality studies; sample size determination; hierarchical binomial models

1. INTRODUCTION

Health services and outcomes research involves understanding access, quality, costs and financing of health care services. Measuring the degree to which health care for individuals is consistent with current knowledge is essential for improving care. To this end, the main objective in quality studies involves comparing the quality of care delivered to cohorts of patients by different health care providers. Providers may be physicians, hospitals, health plans or other meaningful units. Although some health care quality studies may involve
cluster randomization [1, 2] in which clusters are randomized to interventions, the majority are observational studies designed to examine the effectiveness of the delivery of health care, that is, the therapeutic effects of treatment approaches when applied in routine practice.

The challenges in the design and subsequent analyses of observational health quality studies are several-fold. The major issue is that patients are not randomized to providers. This lack of randomization may lead to confounding. Health services researchers have historically adopted one of two approaches to reduce bias in these settings [3]. The first approach involves measuring patient responses that characterize health status, such as survival or health-reported quality of life, and utilizing regression-adjustment techniques to balance the observed confounders across the clusters. The idea is to adjust for the patient’s condition prior to the initial contact with the provider so what remains is the therapeutic effectiveness [4], in addition to random error. For example, 30-day hospital mortality rates after risk-adjustment are often indirectly standardized using the full sample of patients and hospitals, and then compared [5, 6].

The second and currently popular approach involves identifying patients who are expected to benefit from a therapy or process of care and counting how many of these patients actually received the needed therapy. In contrast to the regression-adjustment approach, ‘process-based’ measures are implicitly risk adjusted by restricting the sample to those patients who are known to benefit. For example, rather than determining the rate of beta-blocker utilization in all heart disease patients treated by a physician, researchers examine the use in patients who have no contraindications to beta-blockers [7] and then compare the use among physicians. Only processes of care with a proven positive direct relationship with patient health status are studied so that the higher the rate, the better the quality of care. Eligible patients are identified using established guidelines or other explicit clinical criteria [8, 9]. Because providers are compared using selected subsets of patients, the disadvantages of this approach include achieving a sufficient sample size and generalizing results to all patients treated by a provider.

Another design issue, even after accounting for confounding, is the lack of statistical independence among the observations within a provider [10–12]. Patients treated by the same cardiologist are likely to receive similar treatments, even after risk adjustment, as would patients treated in the same hospital. The statistical dependence inflates the variance of the responses and invalidates statistical tests if not properly accounted for in analyses. In cluster randomization trials, this effect corresponds to the variance inflation factor which is a function of the average cluster size and the intraclass correlation coefficient. Small values of the intraclass correlation coefficient can lead to a substantial inflation on the response variance [13]. In addition to within-cluster dependence, there is also a likelihood of imbalance of cluster characteristics that may influence the response. Stratification on the basis of cluster factors, such as geographic location, may be necessary in order to appropriately model the response.

Lastly, unlike many cluster randomized trials, the target of inference in quality of care studies is the cluster. The inferential goal is to make comparative statements about physicians, medical groups, health plans or hospitals. This requires estimation of cluster-specific parameters as well as functions of the parameters, such as the mean or range of the parameters, across the clusters. The range in surgeon-specific risk-adjusted mortality rates or the average surgeon-specific mortality rate are two examples. Because the estimand of interest is a function of the cluster-specific parameters, often the solutions will be mathematically intractable.

The purpose of this paper is to review approaches to sample size determination to compare providers when designing observational health care quality studies. We focus on process-based
measures because this approach has been widely adopted by many regulatory agencies and health plans [14]. Throughout we use data from a study funded by the Agency for Healthcare Research and Quality designed to develop and test a set of ambulatory quality measures across a continuum of care sites, payment systems and data sources for patients with cardiovascular disease.

The ‘Expansion of Quality Measures for Cardiovascular Disease’ (Q-SPAN-CD) is a two-phase study involving six sites of care corresponding to four major U.S. health plans. Three cohorts were constructed, two hospital-based and one ambulatory-based: patients discharged alive from a hospital with a diagnosis of congestive heart failure (CHF), patients discharged with acute myocardial infarction (AMI), or patients with an outpatient diagnosis of hypertension (HTN). Disease-specific subgroups of clinical experts recommended a set of approximately 30 process-based measures for each disease cohort using explicit clinical criteria [15]. Medical records were abstracted in order to assess conformance with the recommendations. Table I lists the number of eligible patients after a 9-month accrual period for each disease cohort as well as two example process measures per cohort. For example, 396 patients aged 30 years or greater were discharged with a confirmed diagnosis of CHF and met eligibility criteria. Forty per cent of the CHF cohort was eligible for angiotensin converting enzyme (ACE) inhibitors, a medication that reduces blood pressure, increases the heart’s ability to pump blood, and increases the amount of activity the patient is able to do. Patients who are ineligible for ACE include those who have congenital heart disease, those who have had a heart attack within the previous month, and those who have AIDS or kidney failure. In contrast to the CHF cohort, only 3 per cent of the HTN cohort was eligible for ACE inhibitors.

We briefly review sample size requirements in Section 2 in the $K$-cluster setting in which there is dependence among the binary observations within a cluster. Lee and Dubin [16] review methods for sample size determination for correlated binary data and propose an approach to calculating the number of clusters $K$ when estimating $\theta = P(Y_{ij} = 1)$, the per-unit success probability, relying on the central limit theorem. Donner and Klar [17] describe methods for determining the total number of observations as well as the number of clusters from the perspective of cluster randomization trials. Approaches to sample size determination when the investigator directly models the expectation of the response are discussed. A more complete overview of approaches to modelling discrete multivariate data in the presence of correlation can be found in Heagerty and Zeger [18]. Our focus is on hierarchical models (multilevel, cluster-specific or subject-specific models) where the researcher models the expectation of the response conditional on a random effect. In Section 3 we consider stratification in the $K$-cluster setting and characterize sample size when the clusters are not balanced. We also discuss methods when it is reasonable to assume that the underlying rates are exchangeable within a stratum.

2. SAMPLE SIZE FOR DEPENDENT BINARY DATA

Let $Y_{ij} = 1$ denote success for unit $j$ ($j = 1, 2, \ldots, n_i$) in cluster $i$ ($i = 1, 2, \ldots, K$) and 0 otherwise where $n_i$ denotes the total number of units in cluster $i$ and $K$ denotes the total number of clusters. Further, let $\theta_i = P(Y_{ij} = 1)$ represent the success probability for the $j$th unit in the $i$th cluster. For example, $Y_{ij} = 1$ if the $j$th CHF patient treated by the $i$th cardiologist who was eligible for an ACE inhibitor received it and 0 otherwise, $n_i$ denotes the number of patients
Table I. Q-Span-CD Sites. Nine-month accrual rates for all patients meeting study eligibility criteria after initial identification from administrative data. Entries represent the number of patients meeting eligibility criteria for each disease group (cohort size) and for each process measure (number of patients eligible).

<table>
<thead>
<tr>
<th>Site of care</th>
<th>Congestive heart failure (CHF)</th>
<th>Acute myocardial infarction (AMI)</th>
<th>Hypertension (HTN)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort size</td>
<td>Number of patients eligible</td>
<td>ACE Rx</td>
</tr>
<tr>
<td>A</td>
<td>17</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>155</td>
<td>49</td>
<td>53</td>
</tr>
<tr>
<td>C</td>
<td>79</td>
<td>38</td>
<td>41</td>
</tr>
<tr>
<td>D</td>
<td>57</td>
<td>28</td>
<td>30</td>
</tr>
<tr>
<td>E</td>
<td>18</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>F</td>
<td>70</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>396</td>
<td>158</td>
<td>171</td>
</tr>
<tr>
<td>Per cent of cohort</td>
<td>100</td>
<td>40</td>
<td>43</td>
</tr>
</tbody>
</table>

ACE = angiotensin converting enzyme; NSAID = non-steroidal anti-inflammatory drug; Rx = prescription; — site did not contribute patients for the disease group.
treated by cardiologist \( i \) who were eligible for ACE inhibitors, and \( \theta_i \) denotes the probability of receiving the therapy if treated by the \( i \)th cardiologist. We discuss sample size determination using interval-based criteria such as requiring the average width of a \( 100(1 - \alpha) \) per cent confidence interval for functions of \( \theta_i \) to be at most \( w \). In this article our primary interest is in estimation of functions of \( \theta_i \), such as the average cluster-specific success rate or the range in cluster-specific success rates.

### 2.1. Marginal models

There are several approaches available to researchers to analyse clustered binary outcomes, each with specific assumptions regarding the dependence structure of the responses [18]. A standard approach to modelling correlated data is to directly specify the marginal expectation of the individual response separately from the within-cluster correlation [13, 19, 20]. An advantage of these ‘marginal models’ is that they do not require complete specification of the joint distribution of \( Y_i = \{Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\} \). Moreover, these models are robust to misspecification of the dependence structure of the responses.

Donner and colleagues proposed sample size calculations to determine the total number of patients, \( N \), needed in completely randomized cluster designs [21] and in stratiﬁed cluster designs [22] for testing for differences in two treatment groups. In their formulation, the total number of patients needed to estimate a \( 100(1 - \alpha) \) conﬁdence interval for \( \theta_i \) to within a specified width \( w \) is

\[
N = \frac{4Z_{1-\alpha/2}^2 \{\theta(1 - \theta)\} [1 + (n - 1)\rho]}{w^2}
\]

where \( N = nK \) is the total sample size, \( \rho \) generally measures the agreement that exists in a cluster beyond that expected by chance [21], \( n_i = n, \forall i \) and it is assumed that \( \theta = \frac{1}{K}E(\sum_i \hat{\theta}_i) \) where \( \hat{\theta}_i = \frac{1}{n_i} \sum_j Y_{ij} \). When the clusters are of a fixed size, \( \rho \) is interpreted as the average degree of intracluster correlation among cluster members [23] and the design effect, \( \{1 + (n - 1)\rho\} \), is the degree of variance inflation due to clustering averaged among the clusters. When \( \rho = 1 \), the cluster behaves like a single individual while when \( \rho = 0 \), the observations within a cluster are completely independent. For unequal cluster sizes, \( n \) can be replaced by [24] \( \bar{n} = \sum_i n_i^2 / \sum_i n_i \) or, to be conservative, by the largest expected cluster size [21], \( \max(n_i) \).

### 2.2. Hierarchical models

An alternative method of modelling correlation among the responses involves assuming a set of unobserved variables that link the responses. Hierarchical (or multilevel) models [25, 26] separate the within- and between-cluster variability using the following model. Let \( Y_i = \sum_{j}^n Y_{ij} \), the total number of successes in the \( i \)th cluster, and assume

\[
\text{Stage 1: } Y_i | n_i, \theta_i \overset{\text{independent}}{\sim} \text{binomial}(n_i, \theta_i)
\]

In the absence of any other information to distinguish the \( \theta_i \)'s, then symmetry in the prior distribution for \( \theta_i \) is postulated using

\[
\text{Stage 2: } \theta_i | \alpha, \beta \overset{\text{iid}}{\sim} \text{beta}(\alpha, \beta)
\]
where the hyperparameters $\alpha > 0$ and $\beta > 0$ may be prespecified by the investigator. Alternatively, in order to learn about $\alpha$ and $\beta$, distributions for the hyperparameters may be specified using

\begin{equation}
\text{Stage 3: } \alpha \sim \text{gamma}(p_\alpha, q_\alpha) \text{ independent } \beta \sim \text{gamma}(p_\beta, q_\beta); \quad q_i > 0
\end{equation}

where $p_\alpha \geq 1$ and $p_\beta \geq 1$ in order for the posterior distribution to be log-concave.

The main advantage of using hierarchical models is that these models permit estimation of cluster-specific parameters, $\theta_i$, by using information on observations within a cluster as well as information between clusters. The validity of the inferences, however, depends upon the number of clusters [27, 28] as well as correct specification of the distribution for $\theta_i$ in equation (3) [18].

Sample size determination for binary outcomes using the hierarchical model specified by equations (2)–(4) has been proposed from different perspectives. Parmigiani and Berry [29] derived the optimal choice of $K$ and $n$ when interest centres on estimation of $\alpha$ and $\beta$ by maximizing the expected Lindley information [31] for the hyperparameters. Zou and Normand [32] extended the interval-based methods proposed by Joseph and colleagues [33, 34] when interest centres on estimation of functions of the $\theta_i$’s. Cluster size, $n$, or the number of clusters, $K$, is solved by taking expectations of widths of $(1 - \alpha)$ central posterior intervals for the function of the $\theta_i$’s over future realizations of the data. These intervals correspond to the range of values above and below which lies exactly 100($\alpha$/2) per cent of the posterior probability. They used a combination of Markov chain Monte Carlo simulation and parametric approximations to the posterior distribution in the case of balanced cluster sizes, $n$.

2.2.1. Example 1: Avoidance of NSAIDs. An investigator wants to determine the range in health plan specific rates of non-steroidal anti-inflammatory drug (NSAID) use within 6 months of a hospital discharge for patients with CHF. Because some anti-inflammatory medications may cause high blood levels of sodium and potassium which lead to extra body fluid and swollen ankles, chronic use of such medications should be avoided. Based on prior information, an average of 90 per cent of CHF patients within a health plan will not be prescribed NSAIDs. However, many health plans have higher than average avoidance rates, implying that the prior distribution (stage 2) for avoidance of NSAIDs is positively skewed. This prior information can be summarized by postulating that $\theta_i \sim \text{beta}(9, 1)$ (Figure 1, left histogram). Alternatively, the distribution of the cluster parameters may be summarized as $\theta \sim \text{beta}(\alpha, \beta)$ where $\alpha \sim \text{gamma}(9, 1)$ and $\beta \sim \text{gamma}(9, 9)$ (Figure 1, right histogram). The investigator wants to determine the number of patients per health plan, $n$, when interest centres on estimating the range across $K = 25$ plans so that a 95 per cent central posterior interval (CPI) for the range in $\theta_i$ has an average width of $w = 0.20$.

Table II lists the widths of 95 per cent coverage CPIs for $n$ with given $K$ based on 50 Monte Carlo simulations, each of which uses 500 Markov chain Monte Carlo iterates [32]. If the investigator is willing to specify that $\theta_i \sim \text{beta}(9, 1)$, then $n = 50$ patients per plan will be needed to meet the required width of $w = 0.20$ when $K = 25$. However, if the investigator is uncertain about the distribution of the avoidance rates and fits a three-stage model (equations (2)–(4)), then fewer than 50 patients per plan are projected if $K = 25$ (Table II: $w = 0.192$ when $K = 25$ and $n = 50$, three-stage model). The sample size in the latter case can be refined using the methods proposed by Müller and Parmigiani [35]. Simulation methods are used to
Table II. Widths of 95 per cent average coverage central posterior intervals for the range across clusters in avoidance of NSAIDs. Based on 50 Monte Carlo simulations, each using Monte Carlo Markov chain simulations of 500 iterations after a burn-in of 500 iterates.

<table>
<thead>
<tr>
<th>n</th>
<th>Two-stage model: $Y_i \sim \text{binomial}(n, \theta_i)$; $\theta_i \sim \text{beta}(9, 1)$</th>
<th>Three-stage model: $Y_i \sim \text{binomial}(n, \theta_i)$; $\theta_i \sim \text{beta}(\alpha, \beta)$; $\alpha \sim \text{gamma}(9, 1)$; $\beta \sim \text{gamma}(9, 9)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K = 10$</td>
<td>$K = 10$</td>
</tr>
<tr>
<td>10</td>
<td>0.283</td>
<td>0.288</td>
</tr>
<tr>
<td>25</td>
<td>0.234</td>
<td>0.244</td>
</tr>
<tr>
<td>50</td>
<td>0.189</td>
<td>0.192</td>
</tr>
<tr>
<td>75</td>
<td>0.165</td>
<td>0.166</td>
</tr>
<tr>
<td>100</td>
<td>0.146</td>
<td>0.154</td>
</tr>
<tr>
<td>125</td>
<td>0.131</td>
<td>0.137</td>
</tr>
</tbody>
</table>
first determine the sample size across a grid of values and a smoother is applied to the widths at selected sample sizes in order to infer the optimal \( n \).

### 3. STRATIFICATION AND EXCHANGEABILITY

Because historically health care quality has varied by provider characteristics, a common design and analytic strategy involves stratification. For example, variation in hospital-specific
utilization rates often depends upon the presence of an academic affiliation or on the availability of surgical technologies [36]. Other stratification factors involve geographic location, degree of specialization [37] and hospital ownership (for example, for-profit or public). Assume there are \( m_s \) clusters of size \( n_s \) for \( s = 1, 2, \ldots, S \). Clusters are contained within strata and we focus on situations in which there are few strata but many clusters per stratum. The total number of clusters is \( K = \sum_{s=1}^{S} m_s \) and the total number of observations is \( N = \sum_{s=1}^{S} m_s n_s \).

Within each cluster \( i \) contained in stratum \( s \), the distribution of \( Y_{is} \) is binomial with the success probability depending on the stratum in addition to the cluster

\[
Y_{is} \mid n_s, \theta_{is} \overset{\text{independent}}{\sim} \text{binomial}(n_s, \theta_{is})
\]

for \( s = 1, 2, \ldots, S \) strata. Within each stratum \( s \), the success probabilities \( \theta_{is} \) are assumed exchangeable among the clusters such that

\[
\theta_{is} \mid \alpha_s, \beta_s \overset{\text{i.i.d.}}{\sim} \text{beta}(\alpha_s, \beta_s)
\]

Exchangeability implies that within each stratum \( s \), we have no other information to distinguish among the cluster-specific rates, \( \theta_{is} \). To complete the specification, the distribution of the hyperparameters \( \alpha_s > 0 \) and \( \beta_s > 0 \) are postulated to be

\[
\alpha_s \sim \text{gamma}(p_{\alpha_s}, q_{\alpha_s}) \quad \text{and} \quad \beta_s \sim \text{gamma}(p_{\beta_s}, q_{\beta_s}) \quad s = 1, 2, \ldots, S
\]

where \( q_s > 0 \) and \( p_s \geq 1 \). As in the unstratified setting, the required cluster sizes or total number of clusters is determined through direct simulation using software that fits multilevel models such as BUGS [38].

### 3.1. Example 2: stratification on the basis of cluster size

A researcher is interested in examining the average hospital-specific rate of dispensing weight counselling to CHF patients. Although no prior information on conformance is available, the researcher has prior knowledge indicating that hospitals should be stratified by presence of an academic affiliation. Non-teaching hospitals are likely to be located in rural areas, have fewer number of beds, and have less access to invasive cardiovascular technologies. In contrast, teaching hospitals are more likely to be located in urban centres, have more beds, and have more invasive technologies. Based on this information, we set \( S = 2 \) strata. We assume that

(a) the cluster sizes within a particular stratum are the same, for example, \( n_{1s} = n_s \) \( \forall i \in s \) and

(b) the cluster sizes between strata may be unequal, for example, \( n_s = r n_{s'} \) where \( r \) is an integer greater than 0 and \( s \neq s' \).

In order to determine the cluster sizes needed to estimate the mean hospital-specific rate so that the average CPI interval width is within a specified bound, a Monte Carlo simulation is undertaken. The estimand of interest is \( \theta' = \frac{1}{S} \sum_{s=1}^{S} \left\{ \frac{1}{K} \sum_{i \in S} \theta_{is} \right\} \), the arithmetic mean of the hospital-specific rates. Note that other weighting schemes could be used to average the rates [39]. Let \( s = 1 \) denote non-teaching hospitals and \( s = 2 \) denote teaching hospitals. In the absence of prior information, we assume \( \alpha_s \sim \text{gamma}(9, 9) \) and \( \beta_s \sim \text{gamma}(9, 9) \) for \( s = 1, 2 \) so that \( E(\theta_{is}) \approx 0.50 \). Because hospital volume varies by teaching status, with teaching hospitals having more patients, we set \( n_2 = r n_1 \) and \( r = 2.0, 10.0 \) where \( n_1 = 5, 20, 50, 100, 200 \). For example, when \( r = 2.0 \) and \( n_1 = 5 \), then we are assuming that each non-teaching hospital has
$n_1 = 5$ CHF patients eligible to receive weight counselling advice while each teaching hospital has $n_2 = 2 \times 5 = 10$ eligible CHF patients.

Using $K = 10, 20, 50, 100$ clusters, two allocation schemes are used for the distribution of the total number of hospitals between the two strata. Let $m_1$ denote the number of non-teaching hospitals and $m_2$ denote the number of teaching hospitals. In the first scheme, $m_1 = m_2 = 0.50K$ so that there are an equal number of teaching and non-teaching hospitals. For example, if $K = 20, r = 2.0$ and $n_1 = 50$, then stratum 1 comprises 10 non-teaching hospitals each of size $n_1 = 50$ patients, and stratum 2 comprises 10 teaching hospitals each of size $n_2 = 2 \times n_1 = 100$ patients. The total number of patients, $N = 10 \times 50 + 10 \times 100 = 1500$ under this allocation scheme. In the second scheme, more teaching hospitals are assumed by letting $m_1 = 0.20K$ and $m_2 = 0.80K$.

Fifty Monte Carlo simulations are run, fixing $K$ and permitting $n_s$ to vary, for each of the 80 combinations defined by $K(10, 20, 50, 100) \times r(2.0, 10.0) \times n_s(5, 20, 50, 100, 200) \times$ allocation $(m_1 = 0.50K, m_2 = 0.20K)$ and $m_3$. Within each Monte Carlo run, direct simulation from the posterior distribution of $\theta^*$ is used to calculate the average lengths of 95 per cent CPIs for $\theta^*$. Each CPI is based on 500 iterates after an initial burn-in of 500 iterates.

Figure 2 summarizes the results of the simulations. The total sample size ranged from $N = 75$ to 160,000 patients across the 80 combinations. Increasing the sample size, whether through an increase in the number of clusters, $K$, or an increase in the cluster size, $n_s$, results
in smaller interval lengths. For example, under the $m_1 = m_2 = 0.50K$ allocation with $r = 2.0$, a total of $N = 1500$ patients can be generated using $K = 10$ clusters of $n_1 = 100$, $K = 20$ clusters with $n_1 = 50$ or $K = 50$ clusters with $n_1 = 50$. Regardless, the average width of 95 per cent CPIs for $\theta^*$ is 0.042 (Figure 2, upper left corner). If the researcher wishes to estimate the average of the hospital-specific rates so that the average interval length is less than 0.06 and is confined to using $K = 10$ hospitals, the minimum number of patients needed under the equal allocation scheme is $n_1 = 50$ (Figure 2).

3.2. Example 2 Revisited: stratum-specific hyperprior distributions

Suppose that in addition to the information described above, the researcher has reason to believe that conformance will be better in teaching hospitals than in non-teaching hospitals. To reflect this, it is assumed that $\alpha_1 \sim \text{gamma}(9,9), \beta_1 \sim \text{gamma}(9,9)$ as before. However, for teaching hospitals, $\alpha_2 \sim \text{gamma}(9,3)$ and $\beta_2 \sim \text{gamma}(9,9)$, implying $E(\theta_{12}) \approx 0.75$. Rather than the average rate, the researcher wants to estimate the range in hospital-specific rates of dispensing weight advice across all the hospitals to within a specified bound. As before: $K = 10, 20, 50, 100$; $m_1 = m_2 = 0.50K$ or $m_1 = 0.20K$; $m_2 = 0.80K$. The cluster sizes also remain the same with $n_2 = rn_1, r = 2.0, 10.0$ and $n_1 = 5, 20, 50, 100, 200$. Average CPI lengths were computed for the 80 combinations of $K$, $r$, $n$, and $m$, using direct simulation.

Figure 3 displays the widths of 95 per cent CPIs for the range across hospitals in the hospital-specific rates using the various combinations of cluster sizes, stratum allocations and number of clusters. Not surprisingly, the average widths of the range in hospital-specific rates (Figure 3) are more than twice the size of the widths for the mean of the hospital-specific rates (Figure 2). Furthermore, the distribution of patients across the clusters impacts sample size. For example, using a total of $K = 10$ clusters under the $m_1 = 0.20K, m_2 = 0.80K$ allocation scheme, with $m_1 = 2$ clusters of size $n_1 = 100$, $m_2 = 8$ clusters of size $n_2 = 200$ for a total of $N = 1800$ patients, the average width of the 95 per cent CPIs is 0.130. However, using a total of $K = 20$ clusters with $m_1 = 4$ clusters of size $n_1 = 50$ and $m_2 = 16$ clusters of size $n_2 = 100$ for a total of $N = 1800$ patients, the average width of 95 per cent CPI for the range in hospital-specific rates is 0.140. Thus, with a fixed total number of patients, $N$, fewer total clusters $K$ would be needed. For small cluster sizes ($n_1 = 5$), the more extreme the cluster allocations across the different strata, the wider the CPI (Figure 3, upper and lower left column).

4. CONCLUDING REMARKS

In this paper we provided an overview of approaches for sample size determination in observational health care quality studies when interest centres on making comparative inferences about the clusters. Our focus was on binary outcomes using hierarchical models. Although marginal models have the advantage of not requiring any distributional assumptions and provide consistent estimates of standard errors even if the wrong dependence structure is used, they do not provide cluster-specific estimates of parameters. Unlike marginal models, hierarchical models permit researchers to separate between- and within-cluster variability, thereby providing inferences for individual clusters [30].

There have been few general guidelines for sample size requirements in hierarchical designs despite the considerable published research regarding the analyses of multilevel data.
Cohen [40] derived optimal sample size in the hierarchical linear model by minimizing the asymptotic variance of the maximum likelihood estimator of the variance components but the focus was on inference for the regression coefficients. Parmigiani and Berry [29] described a decision theoretic approach when interest centres on estimation of the beta parameters characterizing the distribution of the success probabilities when the outcome is binary in a balanced design. Zou and Normand [32] described interval-based methods for sample size determination for functions of the cluster parameters in balanced binomial hierarchical designs. A general methodology for sample size determination in complex settings that does not rely on asymptotics is that proposed by Müller and Parmigiani [35] where stochastic optimization methods through curve fitting of Monte Carlo samples are employed.

In general, health care quality studies often include several process-based measures for a cohort. We described sample size requirements for functions of the cluster-specific effects using a single binomial outcome per cluster. A reasonable assumption is that the multiple outcomes measured on a provider are related because each reflects the underlying quality of the health care provider. An integrated analysis that combines information across the sets of measures within a provider would provide a more comprehensive analysis [41] and perhaps reduce the total number of patients needed. Sample size in this setting could be determined using stochastic optimization methods [35].
Based on the phase I results of the Q-SPAN-CD study, patient accrual using process measures for patients with cardiovascular disease resulted in a lower number of patients than anticipated. These low numbers arose mostly from a lack of diagnostic test information such as ejection fraction. For example, only 158 CHF patients across six health plans over a 9 month period were eligible for ACE inhibitors while only 68 AMI patients were eligible for cholesterol-lowering medications. When researchers use patient outcomes, such as survival, to compare providers, they will usually retain more patients but will need to develop and validate appropriate regression adjustments. Process-based measures on the other hand, are thought to be actionable [3] and have more face validity, but are only applicable to a selected subset of the study population. Health care quality studies of disease-based cohorts will need to accrue patients over a longer period of time, model multiple outcomes within a provider, or combine measures, in order to enroll sufficient numbers of patients to make valid comparative inferences.

In conclusion, many organizations release quality information for public and regulatory consumption. The National Committee for Quality Assurance publicly releases information on a set of performance measures, known as the Health Employer Data and Information Set, on nearly 300 health plans. The Joint Commission on Accreditation of Health Care Organizations evaluates approximately 20,000 health care organizations, including more than 5000 hospitals. In New York State, surgeon-specific in-hospital mortality for patients undergoing isolated coronary artery bypass graft surgery have been publicly reported since 1989. Between 1991 and 1995, information on more than 180 surgeons for approximately 70,000 patients was available. While the degree of statistical dependence at the health plan level may be minimal, the same is not true at the hospital and surgeon levels. Hierarchical models should be employed when the focus of inference is to compare clusters. Furthermore, because researchers will want to learn about performance across providers, a three-stage model which permits learning about the hyperparameters is preferred to a two-stage model. A particularly important contribution by statisticians who work in the area of health care quality will involve working with health services researchers to emphasize the importance of appropriate methodology, and to develop needed methodology and software.

ACKNOWLEDGEMENTS

We thank Professor K. C. Carrière, University of Alberta, Edmonton, for organizing this conference and inviting us to present our work. We also thank Dr Barbara J. McNeil for helpful comments on earlier drafts and Ms Mariko Golden for word processing assistance, both from Harvard Medical School, Boston, U.S.A. We are also indebted to two anonymous referees for valuable suggestions to improve the clarity of the article.

REFERENCES


