ORIGINAL RESEARCH

Can Choice of the Sample Population Affect Perceived Performance: Implications for Performance Assessment

Bruce E. Landon, MD, MBA[1,2], A. James O'Malley, PhD[1], and Thomas Keegan, PhD[1]

[1]Department of Health Care Policy, Harvard Medical School, Boston, MA, USA; [2]Division of General Medicine and Primary Care, Beth Israel Deaconess Medical Center, Boston, MA, USA.

BACKGROUND: There is accelerating interest in measuring and reporting the quality of care delivered by health care providers and organizations, but methods for defining the patient panels for which they are held accountable are not well defined.

OBJECTIVES: To examine the potential impact of using alternative algorithms to define accountable patient populations for performance assessment.

RESEARCH DESIGN: We used administrative data regarding Community Health Center (CHC) visits in simulations of performance assessment for breast, cervical, and colorectal cancer screening.

PARTICIPANTS: Fifteen CHC sites in the northeastern US.

MEASURES: We used three different algorithms to define patient populations eligible for measurement of cancer screening rates and simulated center-level performance rates based on these alternative population definitions.

RESULTS: Focusing on breast cancer screening, the percentage of women aged 51–75 eligible for this measure across CHCs, if using the most stringent algorithm (requiring a visit in the assessment year plus at least one visit in the 2 years prior), ranged from 28% to 60%. Analogous ranges for cervical and colorectal cancer screening were 18–59% and 26–62%, respectively. Simulated performance data from the centers demonstrate that variations in eligible patient populations across health centers could lead to the appearance of large differences in health center performance or differences in expected rankings of CHCs when no such differences exist. For instance, when holding performance among similar populations constant, but varying the proportion of populations seen across different health centers, simulated health center adherence to screening guidelines varied by over 15% even though actual adherence for similar populations did not differ.

CONCLUSIONS: Quality measurement systems, such as those being used in pay-for-performance and public reporting programs, must consider the definitions used to identify sample populations and how such populations might differ across providers, clinical practice groups, and provider systems.

Received April 30, 2009
Revised September 8, 2009
Accepted September 26, 2009
Published online November 21, 2009

J Gen Intern Med 25(2):104–9
DOI: 10.1007/s11606-009-1153-z

INTRODUCTION

Since a series of influential reports released by the IOM, there has been accelerating interest in measuring and improving the quality of care in the US health care system[1,2]. More recently, efforts such as pay-for-performance and public reporting have raised the stakes for physicians and physician groups[3]. If such quality measurement and improvement programs are to be successful, the performance data used to assess them must accurately reflect the quality of care delivered so that any discrepancies with recommended care do not simply reflect faulty measurement or flawed attribution.

One frequently heard complaint from physicians is that the patient panels for which they are held accountable, whether developed by health plans or even by their own practice information systems, are often inaccurate. Generally, this is due to the inclusion of patients who have left their practice, were never seen in their practice, or with whom they had a transient relationship. Including such patients in physician profiles undermines the profiling system in two ways. First, the profiles might be inaccurate. To the extent that such profiles are used to determine compensation or to inform patient choice, there could be serious ramifications for physicians' practices and patients. Second, such inaccuracies engender mistrust of the system by those being measured. Physicians become concerned both that their performance data are inaccurate and that other aspects of the profiling system are similarly error prone. Since physician acceptance of performance assessments is important for changing practice methods, confidence in these assessments is crucial. Data errors shift attention away from improving performance towards improving documentation and the accuracy of the data, and thus undermine the usefulness of such data for improving quality[4].

A first step in assessing performance is assuring that patient panels of individual physicians or physician practices comprise the patients to whom they are actually providing care. Although patients in some HMO-type health plans usually have to identify a primary care physician, most patients, including those in traditional Medicare and Preferred Provider-type health plans, do not. Outside of such explicit

arrangements, the most accurate way of ascertaining such panels is to directly ask patients and physicians, but such a process would be time consuming and expensive, so administrative data often are used for this purpose. Using administrative data presents challenges, however, because no combination of visits and duration of contact perfectly captures the true patient population. Some algorithms might be overly stringent (thus lacking sensitivity), whereas others might be less stringent (lacking specificity). Determining patient panels for Community Health Centers (CHCs) adds an additional layer of complexity because of the high proportion of under- or uninsured patients, whose care cannot be identified through insurer data systems. Such panels can only be ascertained through visit history or registration information kept at the CHC.

Many observers dismiss these types of potential pitfalls, arguing that all are subject to the same types of errors and that these errors are distributed randomly across physicians and organizations. Thus, it is claimed they do not impact assessments of relative performance. To date, however, no one has examined whether this assumption is true. In this paper, we use empirical data collected from nine CHCs in the northeastern US on visits to health centers to parameterize and quantify the potential impact of using alternative algorithms to define patient populations for performance assessment. In particular, we sought to determine if the choice of sample population could potentially influence judgments about the performance of individual CHCs.

METHODS

Overview

We obtained administrative data on patient visits from CHCs located in the Northeast US (see Table 1). We then used audited medical record abstraction data to estimate overall cancer screening rates at these health centers and then performed simulations to estimate the effect of using different sample populations on observed screening rates. We assigned hypothetical patients cared for at each CHC to three mutually exclusive groups, which we defined based on the frequency and duration of contact with the health center. These mutually exclusive groups were then used either alone or in combination to define three different sample populations. In the base case scenario, screening rates vary across the groups (e.g., those that had only been seen in the center once in the past 3 years were assumed to have lower screening rates than those seen more frequently) but are identical across CHCs. If the proportion of patients in each of the groups varied by center, however, center-level performance assessments and rankings based on different sample populations could vary based on the groups used to define the sample population for the measure. In order to perform the simulations, we let screening rates vary according to the level of continuity of visits over the past 3 years. We focus on screening measures for breast, cervical, and colorectal cancer. Of note, the medical record data only yield overall screening rates. Due to sample size limitations, we were not able to empirically estimate screening rates for each of these sub-populations.

Table 1. Community Health Center Characteristics

Center	Location	No. primary care sites	Total patients in 2004[a]
A	CT	6	39,138
B	CT	4	26,406
C	NY	7	23,015
D	NY	5	27,622
E	NY	3	25,850
F	NY	3	21,281
G	MA	2	8,199
H	MA	1	7,855
I	MA	1	7,990
Mean		3.6	20,817

[a]*Obtained from the HRSA Uniform Data System*

Study Population

We obtained administrative data on visit history for the years 2002–2004 from 15 sites at nine CHCs in the Northeast US taking part in a pilot study of cancer screening. The only requirement for participating in the study was that the centers needed to provide an anonymized electronic list of all patients seen during a 3-year time period at up to two of their care sites. For each patient, we also obtained age and sex, in order to determine measure eligibility, and information on whether they had any encounters during the assessment year (calendar year 2004), or during the preceding two calendar years.

Patient Population Identification Algorithms

We used the patient visit data to define potential quality assessment populations for each health center during the assessment year 2004. We first defined target populations eligible for breast cancer screening (women aged 41–75), colorectal cancer screening (men and women aged 50–75), and cervical cancer screening (women aged 21–75) based on guidelines from the USPSTF and NCQA specifications[5–7]. We then estimated screening rates based on sample populations drawn using three sampling algorithms that varied by degree of inclusiveness: aample design A, any visit during the assessment year *or* the 2 years prior; sample design B, any visit during the assessment year; sample design C, any visit during the assessment year *and* at least one visit during the 2 years prior to the assessment year. Sample design A defines the most inclusive patient population and includes all patients with at least one visit over the 3-year time period. While it may seem overly inclusive, this definition is most consistent with the community-oriented mission of CHCs. Sample design C is the least inclusive, and requires some degree of continuity because the patient had to have been seen during the assessment year and at least once in the prior 2 years. Sample design B includes all patients seen at least once during the assessment year, some of whom might have been seen just a single time for an episodic condition and who might have had no history of routine care at the center. For reporting purposes to HRSA, the Uniform Data Set (required annually by HRSA) defines patient populations according to patient encounters in the past year as in sample design B.

There is some overlap in the patients who would fall into the above sample designs (e.g., a patient seen in 2004 and at least once in the prior 2 years would be included using any of these three designs). In order to understand the consequences of using these various sample designs, we decomposed them into

three mutually exclusive groups of patients: those seen in either 2002 and/or 2003 without having had a visit in 2004 (group one), those seen in 2004 only (group two), and those seen in 2004 *and* either 2002 or 2003 (group three). We then used these groups alone or in combination to construct each of the assessment populations according to the sample designs described above. The mutually exclusive patient groups and how these are combined by the sample designs are depicted in Figure 1.

Analyses

Using the administrative data supplied by each of the health centers, we first describe and quantify the distribution of patients qualifying for each of the three screening tests under each sample design (A, B, or C). We then examine the influence on center-level performance of using different sample design algorithms to define the eligible denominator population for each screening test. We use overall rates of cancer screening observed in our pilot study of CHCs as a basis for setting a priori screening rates under four different scenarios in the simulation. The screening rates associated with each of these groups were also varied across simulations.

Simulations using a range of plausible screening rates were used because we could not obtain sufficient numbers of patients in the chart review study to estimate actual screening rates for the exclusive groups of patients defined above. The observed screening rates were based on chart audits of random samples of 40 eligible patients per screening test from eight of the CHCs. These sample-based screening rates, representing the range of rates that might plausibly be observed in the CHC population, were then used as boundaries for the set of screening rates used in our simulations.

We first assume that performance (screening rates) is equal across health centers for each of the three mutually exclusive groups used to define the universe of potentially eligible patients under each of the sample designs, which implies that any differences seen in center-level screening rates arise because of differences in the distribution of patients from each of these groups across the centers. In addition, our starting case is one where the distribution of the groups across the nine centers is also identical. In this scenario, the choice of sample design algorithms used to define the denominator population has no effect because there are no differences in performance between the eligible groups used to define the denominator population and the distribution of the groups are identical across the centers.

We then vary the screening rates across the three mutually exclusive groups based on observed variations within our nine health centers. We assign the highest screening rate to patients in group three (those with some evidence of continuity) and the lowest to group one (which includes patients who only have been seen at the center on a single occasion 1 or 2 years prior to the assessment year). By definition, some patients in the latter group would not be up to date in screening because a test would have been required during 2004 (e.g., breast cancer screening), when they were not seen at the center. The assigned rate for group two (those seen only during the assessment year) falls in the middle between these extremes.

We then calculate expected screening rates for each of the potential sample designs after varying the composition of the health center population based on the empirical data from the nine centers on the actual proportion of patients that would fall into each of the groups. For example, for CHCs with a high degree of continuity (e.g., most patients are in group three, those seen at least once in 2004 and at least once in either 2002 or 2003), there will be less difference in performance when the sample design is varied because most patients are from group three, and group three is included in each sample design. However, for centers with more patients from the other populations (indicating less continuity), there could be a larger effect. As noted earlier, we used the empirical data on overall cancer rates from the health centers to provide natural boundaries of the distributions.

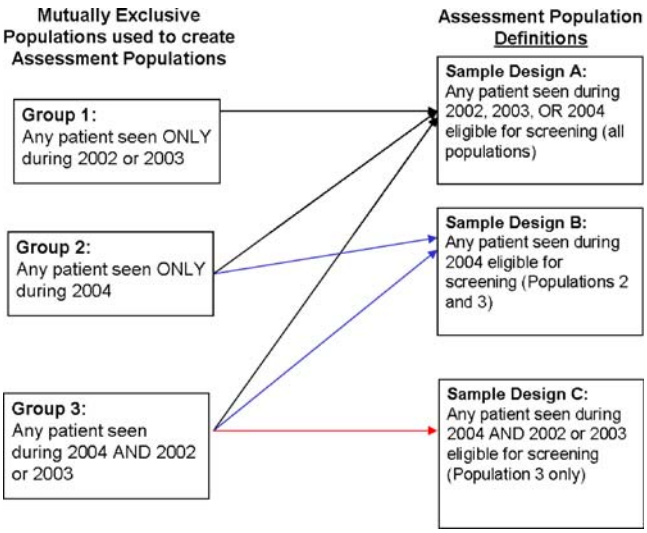

Figure 1. Schematic of the population groups comprising sample designs A, B, and C.

RESULTS

Community Health Center and Study Population

Characteristics of the CHCs are presented in Table 1. Two of the health centers were located in Connecticut, four in New York, and three in Massachusetts. The centers cared for approximately 20,000 patients each, with an average of eight care delivery sites at each center.

Table 2 presents the distribution of eligible patients for each of the cancer screening tests at the delivery site level when using the three sample designs to define the denominator populations. As noted above, sample design C is the least inclusive, requiring a visit during the assessment year and at least one in the 2 years prior, whereas sample design A uses the most inclusive criteria, including all patients with at least one visit during the 3-year time period.

For illustrative purposes, we focus on the results for breast cancer screening and highlight two care sites at the extremes of the distribution. Site C1 had a total of 695 eligible women that had at least one visit over the 3-year period (sample design A). Of these, however, only 193 (28%) met the criteria for sample design C that requires at least one visit in the assessment year

Table 2. Distribution of Eligible Patient Populations for Screening According to Sample Design (SD) Definition

	Breast cancer screening				Cervical cancer screening				Colorectal cancer screening			
Center	SD A	SD B	SD C	SD C as a % of total	SD A	SD B	SD C	SD C as a % of total	SD A	SD B	SD C	SD C as a % of total
A1	1,140	741	589	52	2,802	1,646	1,211	43	1,011	654	535	53
A2	1,526	1,081	887	58	2,939	1,860	1,457	50	1,512	1,046	866	57
B1	2,832	1,763	1,257	44	5,862	3,224	2,046	35	2,821	1,801	1327	47
B2	682	451	295	43	1,431	884	529	37	541	365	236	44
C1	695	386	193	**28**	1,774	931	391	22	547	336	164	30
C2	1,470	929	438	30	4,445	2,413	819	**18**	1,379	877	459	33
D1	297	179	97	33	628	399	185	30	265	167	85	32
D2	3,475	2,554	1,630	47	7,441	5,341	3,180	43	2,897	2,112	1,318	46
E1	404	255	190	47	1,102	645	424	39	310	212	155	50
E2	4,508	3,034	2,305	51	8,170	5,089	3,491	43	4,339	2,936	2,202	51
F1	3,689	2,480	2,181	59	10,903	7,493	6,357	58	3,075	2,022	1,773	58
F2	2,266	1,628	1,357	60	6,851	4,988	4,018	**59**	1,824	1,278	1,088	60
G	1,977	661	576	29	5,101	1,727	1,436	28	1,768	516	455	**26**
H	1,042	733	623	60	2,913	1,861	1,360	47	1,252	903	776	**62**
I	808	686	501	**62**	1,967	1,637	1,092	56	685	589	398	58
Total	26,811	17,561	13,119	49	64,329	40,138	27,996	44	24,226	15,814	1,1837	49

Sample design A. All patients seen at any time during 2002–2004
Sample design B. All patients with at least one visit in 2004
Sample design C. All patients with at least one visit in 2004 AND at least one visit in 2002 or 2003
The bold numbers indicate the lowest and highest percentages across the CHCs

and at least one visit in the 2 years prior. In contrast, for Center I, 62% (501 patients) of the 808 eligible patients seen over the 3-year time period met the criteria required for sample design C. Similar results are seen for the cervical cancer and colorectal cancer screening populations.

Simulated Effects on Measured Screening Rates

In Table 3, we present the range of results that might be expected if the actual screening rates for the three mutually exclusive groups are identical across health centers, but we

Table 3. Estimated Screening Rates When Distribution of Patient's by Group and Screening Rates for Patients in Specific Groups Differ Across Centers

	Population distribution by group			Expected performance (overall screening rate)		
	Proportion of CHC patients in group 1	Proportion of CHC patients in group 2	Proportion of CHC patients in group 3	Sample design A (most inclusive)	Sample design B	Sample design C (least inclusive)
	Actual screening rates by patient group			(Sampled from groups 1, 2, 3)	(Sampled from groups 1, 3)	(Sampled from group 3)
Scenario 1. Screening rates are identical across groups	1=0.5	2=0.5	3=0.5			
	0.33	0.33	0.33	0.50	0.50	0.50
	0.35	0.50	0.15	0.50	0.50	0.50
	0.30	0.35	0.35	0.50	0.50	0.50
	0.20	0.25	0.55	0.50	0.50	0.50
	0.10	0.15	0.75	0.50	0.50	0.50
Scenario 2: Screening rates vary minimally across groups	1=0.45	2=0.5	3=0.55			
	0.33	0.33	0.33	0.50	0.525	0.55
	0.35	0.50	0.15	0.49	0.512	0.55
	0.30	0.35	0.35	0.503	0.525	0.55
	0.20	0.25	0.55	0.518	0.534	0.55
	0.10	0.15	0.75	0.533	0.542	0.55
Scenario 3: Screening rates vary moderately across groups	1=0.35	2=0.5	3=0.65			
	0.33	0.33	0.33	0.50	0.575	0.65
	0.35	0.50	0.15	0.407	0.535	0.65
	0.30	0.35	0.35	0.508	0.575	0.65
	0.20	0.25	0.55	0.553	0.603	0.65
	0.10	0.15	0.75	0.598	0.625	0.65
Scenario 4: Screening rates vary greatly across groups	1=0.25	2=0.5	3=0.75			
	0.33	0.33	0.33	0.50	0.625	0.75
	0.35	0.50	0.15	0.45	0.558	0.75
	0.30	0.35	0.35	0.513	0.625	0.75
	0.20	0.25	0.55	0.588	0.672	0.75
	0.10	0.15	0.75	0.663	0.708	0.75

Standard deviations are all approximately 0.05 so these are not presented individually
Estimates are derived for sample designs A, B, and C. Each sample design encompasses a random sample of patients from three mutually exclusive groups (sampled according to their distribution at the center) (see Fig. 1)
Group 1 patients were seen only in 2002 or 2003. Group 2 patients were seen only in 2004 (the assessment year). Group 3 patients were seen in 2004 AND at least once in either 2002 or 2003

vary the screening rates across these groups and the distribution of the groups within a health center. For each scenario, we begin with the case where the groups are evenly distributed (i.e., each group composes 33% of the clinic population) and then vary the population distribution that is sampled from under the various sample designs, with the degree of variation roughly approximating the most extreme distribution observed in the empirical data (see Table 2). In the first set of rows in Table 3 (scenario 1), we begin with the case where the actual screening rates of the groups are identical (rate = 50%) so we observe no change in the expected screening rates.

Scenarios 2–4 are identical in procedure, with the exception that we vary the differences in screening rates for the three mutually exclusive groups by increasingly large amounts (ranging from 10 percentage points to 50 percentage points between the highest and lowest performing groups). Notably, within each scenario, the rates for sample design C are always identical because even though the proportion of patients from that sample population varies across centers, this sample design includes only patients from group three. In scenario 2, when the difference in rates between group three (with the highest performance) and group one (with the lowest performance) is set at 10 percentage points, the observed differences in center level performance when using sample designs A and B are small at 4.3% for sample design A (range = 49.0% to 53.3%) and 3% for sample design B (range = 51.2% to 54.2%). As explained above, there is no variation for sample design C.

The differences in expected performance for the third and fourth scenarios are progressively more pronounced. In the fourth scenario, the screening rates for the three populations vary from a low of 25% for population one to a high of 75% for population three. Differences in observed screening rates among centers when using sample design A might vary from a low of 45% (for centers with relatively few patients in group 3) to 66.3% (where most patients are in group 3), a difference of more than 20 percentage points. Similar results are observed for sample design B, ranging from 55.8% (in the second set of rows) to 70.8% (in the final set of rows).

DISCUSSION

Pay-for-performance systems and public reporting of performance data are becoming increasingly prevalent in our health care system[3]. Although mostly overlooked in discussions of this topic, identifying the appropriate patient population for performance assessment is crucial to the design and implementation of these systems because using various methods of identifying patient populations might lead to large differences in observed performance that are, in fact, spurious. In this study, we present data that demonstrate the potential impact of different sample design decisions on health center assessments of cancer screening rates when the populations vary across centers. This approach has potentially broad implications because variation of the type we observed in our health centers and simulated in our analyses are broadly applicable within primary care.

We find several notable results. First, regardless of sample design, we find large numbers of individuals eligible for each of the screening tests we examined at all of the health centers. Second, we show wide variation in the proportion of health center patients that would be included in the denominator population, with almost three-fold variation in the proportion of health center patients that would be eligible for the denominator using the most divergent sampling criteria. This implies that fairly subtle differences in selection criteria may lead to very different assessment groups in some settings. Finally, using simulation techniques, we also show that variations in patient populations reflecting different levels of continuity of care across health centers could lead to the observation of large differences in health center performance even when actual performance rates are identical for the populations of interest, as in the scenario we designed.

The sample design chosen for performance assessment should vary according to the purposes of the assessment[4]. For assessment systems that are going to be used for external purposes, including public release of performance information or pay-for-performance, assessment systems should strive to minimize differences across physicians (or health centers in this case) that arise due to differences in the patient populations served, as opposed to underlying performance. Alternatively, if assessment is being used primarily to track performance or for internal quality improvement efforts, one might desire a more inclusive definition that would afford centers the possibility of identifying populations of patients that might benefit from more intensive outreach or other efforts.

Our study also highlights the importance of choosing the appropriate denominator population in order to restore confidence in the validity of performance assessments. Physicians may lose faith in performance assessments if the patient selection criteria do not match their own perceptions of the patients for whom they are responsible. The present results, by showing that different sample designs could lead to different assessments, suggest that denominator populations should be chosen carefully based on the target audience of the assessment results. For example, if a goal is to change physician behavior, we may need to choose a sample of patients who match physicians' own definition of patients to whom they are responsible.

Our study is subject to several important limitations. First, our empirical data collection was limited to administrative data from 15 sites at 9 health centers located in a single region of the country. However, the purpose of these data is to illustrate the potential variation across health centers and, even with this small sample size, we saw important variation in the proportions of patients that would qualify under various sample design scenarios. A bigger limitation is the lack of empirical data on real differences that might be observed among these populations. Although the overall estimated rates we used were consistent with prior data, we were not able to sample sufficient numbers of patients from each of the mutually exclusive groups to enable precise estimates of this portion of variation. Nonetheless, we believe that our simulated results represent a reasonable range of results that might be observed under these sample designs.

In summary, we find that performance assessments of community health centers are likely to be sensitive to the sample designs used to define the patient populations for whom they are responsible. The potential differences we observe are clinically important and could have important ramifications for CHCs if connected to funding or public reporting. Moreover, our results also have implications for public reporting and pay-for-performance programs that face similar methodological challenges. Our results suggest that assessment systems must carefully consider the definitions

used to identify accountable populations and how such populations might differ across health care settings.

Acknowledgements: *This project was supported by grant numbers 1R01CA112367–01 from the National Cancer Institute and 1 U01 HS13653 from the Agency for Healthcare Research and Quality, with support from the Health Resources and Services Administration.*
The authors thank Yang Xu, M.S., for statistical programming, Emily Corcoran for editorial assistance, and Steve Taplin, M.D., for comments on an earlier version of this manuscript. We also thank the participating Community Health Centers without whom this research could not have been completed.

Conflict of Interest: *None disclosed.*

Corresponding Author: *Bruce E. Landon, MD, MBA; Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02215, USA (e-mail: landon@hcp.med.harvard.edu).*

REFERENCES

1. Institute of Medicine. To Err is Human: Building a Safer Health System. Washington, DC: National Academy Press; 2000.
2. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academy Press; 2001.
3. **Rosenthal M, Landon BE, Normand SL, Frank RG, Epstein AM.** Pay for performance in commercial HMOs. N Engl J Med. 2006;355:1895–1902.
4. **Landon BE, Normand ST, Blumenthal D, Daley J.** Physician clinical performance assessment: prospects and barriers. JAMA. 2003;290: 1183–9.
5. Agency for Healthcare Research and Quality. Screening for Breast Cancer: Summary of Recommendations/ Supporting Documents, Available at: http://www.ahrq.gov/clinic/uspstf/uspsbrca.htm, Accessed October 10, 2009.
6. Agency for Healthcare Research and Quality, Screening for Colorectal Cancer, Available at: http://www.ahrq.gov/clinic/uspstf/uspscolo.htm, Accessed October 10, 2009.
7. Agency for Healthcare Research and Quality, Screening for Cervical Cancer, Available at: http://www.ahrq.gov/clinic/USpstf/uspscerv.htm, Accessed October 10, 2009.