Measuring Health Care Quality and Value: Theory and Empirics

Measuring Health Care Quality and Value: Theory and Empirics

A dissertation presented

by

Aaron Lawrence Schwartz

to

The Committee on Higher Degrees in Health Policy
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Health Policy

Harvard University
Cambridge, Massachusetts

April, 2015

Dissertation Advisor: Joseph P. Newhouse, Ph.D. Aaron Lawrence Schwartz

Measuring Health Care Quality and Value: Theory and Empirics

Abstract

Imperfect information is a pervasive feature of health care markets. Therefore, measuring the quality and value of health care services may inform efforts to improve health care delivery. This dissertation explores several applications of performance measurement in health care: describing national practice patterns, evaluating the effects of payment reforms, and contributing to policies that reward providers for measured performance.

Chapter one describes the use of low-value services in fee-for-service Medicare. Drawing from evidence-based lists of services that provide minimal clinical benefit, I develop 26 claims-based measures of low-value services. Applying these measures to Medicare claims, I demonstrate that 42% of beneficiaries received at least one of these services in a year, which constituted 2.7 % of overall annual spending. When more specific and less sensitive versions of the measures were used, I detected low-value service use for 25% of beneficiaries, constituting 0.6% of overall spending. In adjusted analyses, spending on low-value services was substantial even in regions at the 5th percentile of the regional distribution of low-value spending. Adjusted regional use was positively correlated among five of six categories of low-value services. These findings are consistent with the view that wasteful practices are pervasive in the US health care system. The results also

suggest that the performance of claims-based measures in supporting policies to reduce overuse may depend heavily on how the measures are defined.

Chapter two examines the role of provider organizations in influencing the delivery of low-value services. In Part I of this chapter, I assess whether provider organizations exhibit distinct profiles of low-value service use in fee-for-service Medicare. In one sample of 3,137 large provider organizations and another sample of 250 provider organizations that entered the Medicare Pioneer Accountable Care Organization (ACO) Program or the Medicare Shared Savings Program, I demonstrate that provider organizations' use of low-value services exhibits considerable variation, substantial persistence over time, and modest consistency across service types. In Part II of this chapter, I evaluate the effects of the Pioneer ACO Program on the use of low-value services. In a difference-in-differences analysis, I compare the use of low-value services between beneficiaries attributed to Pioneer ACOs and beneficiaries attributed to other providers, before (2009-2011) vs. after (2012) Pioneer ACO contracts began. During its first year, the Pioneer ACO program was associated with modest reductions in low-value services, with greater reductions for organizations that had provided more low-value services. The findings in this chapter suggest that provider organizations can influence the use of low-value services by affiliated physicians, and that organization-level incentives can reduce low-value practices.

Chapter three analyzes the economic properties of performance measures used in both health care and education policy. Because observable outcomes constitute a noisy signal of performance in these settings, shrinkage estimators are

often used to improve measurement accuracy. I demonstrate that these improvements in accuracy come at the cost of reducing a measure's responsiveness to agent behavior, thereby diluting incentives for performance improvement. In a model of consumers sorting between agents, I show that welfare depends on two components: (1) accuracy of performance signals, which promotes efficient consumer sorting, and (2) incentives for performance improvement, which promote efficient agent effort. Using Monte Carlo simulation, I evaluate the accuracy and incentive properties of various techniques for estimating hospital performance in heart attack mortality. Shrinkage estimators entail substantial incentive distortions, particularly for smaller hospitals, which experience an approximate 50-70% "tax" on improvement. Several estimation techniques, including the methods currently used by Medicare, are dominated on the basis of both accuracy and incentive criteria. I discuss various policy alternatives to shrinkage estimation, such as increasing the timespan of measuring performance.

Table of Contents

Acknowledgements

This research would not have been possible without the outstanding contributions of my dissertation committee. Joseph Newhouse drew from his seemingly infinite fund of knowledge about health economics to provide me with invaluable advice throughout my graduate studies. Michael Chernew taught me the importance of working hard to gain traction on challenging but rewarding research topics. Michael McWilliams patiently and assiduously guided me through all manner of obstacles arising during my research, exhibiting an extraordinary commitment to precision in thought and in word. Moreover, I was extremely fortunate to have such kind and generous mentors. Our meetings often brightened my days and inspired me to reach for new questions and answers. Joe, Mike and Michael– I am so grateful for the time and energy you invested into training me, and for serving as role models for the type of scholar I hope to become.

Other faculty also served as excellent advisors, collaborators, and mentors. I benefited enormously from years of discussions with Thomas McGuire, who has a remarkable ability to distill a vague early research idea into its essence. I am indebted to Anupam Jena and Benjamin Sommers for demonstrating how to pursue rewarding careers spanning the worlds of medicine, policy, and economics. Many others in the broader Harvard health policy community provided me with valuable research advice, including: Chris Afendulis, Katherine Baicker, Amitabh Chandra, David Cutler, David Grabowski, John Hsu, Haiden Huskamp, Mary Beth Landrum, Ateev Mehrotra, Laura Hatfield, Frank Levy, Barbara McNeil, Meredith Rosenthal, and Kathy Swartz. I am particularly thankful to faculty who served as coauthors of

portions of this dissertation: Michael Chernew, Bruce Landon and Michael McWilliams (Chapters 1 & 2), Adam Elshaug (Chapter 1), and Alan Zaslavsky (Chapter 2 Section 1).

I shudder to think what graduate school would have been without the camaraderie of fellow travelers. I am especially thankful for the intellectual and emotional support provided by the following health economists in training: Martin Andersen, Sebastian Bauhoff, Abby Friedman, Tim Layton, Dan Ly, Hannah Neprash, Daria Pelech, Sam Richardson, Adam Sacarny, Tisamarie Sherry, Julie Shi, Zirui Song, Ariel Stern, and Jacob Wallace. I hope that graduate school was just the first step in careers full of sharing ideas and good times together.

Generous funding was provided by the National Science Foundation Graduate Fellowship, a National Institute on Aging Predoctoral M.D./Ph.D. National Research Service Award (F30 AG044106-01A1), an Agency for Healthcare Research & Quality (AHRQ) Graduate Fellowship (2T32HS000055-20), and the Harvard Medical School M.D./Ph.D. Program.

Special thanks go to the wonderful Health Policy PhD Program staff members, especially Debbie Whitney and Joan Curhan, who built a strong community for graduate students. Thanks also to the coordinators of the M.D./Ph.D. Program, Anne Becker, Amy Cohen, and Loren Walensky, for their devotion to training physician-investigators in the social sciences. My dissertation research took place at the Harvard Medical School Department of Health Care Policy and the National Bureau of Economic Research, where many staff helped me to pursue my

work, including Emily Corcoran, Jesse Dalton, Ayan Elmi, Elizabeth Haak, Pasha Hamed, Jean Roth, and Katya Zelevinsky.

Most of all, thanks to my family, whose guidance and support has been instrumental to my work and essential to my well-being. Ben and Evelyn – thank you for all of your encouragement even if you did not know quite what I was up to. Katherine – you have been there for me every day, cheering me on in endlessly creative ways, inspiring me with your hard work, and growing with me. Thank you for being my constant source of strength and joy. I love you.

Finally, Mom and Dad – you were my first teachers and your devotion to that role has not wavered over decades. You nurtured my love of learning and taught me persistence by example. For all you have done to make me the person I am today, I dedicate this dissertation to you.

Chapter 1

Measuring Low-Value Care in Medicare*

* A version of this chapter was previously published:

Schwartz AL, Landon BE, Elshaug AG, Chernew ME, McWilliams JM. Measuring low-value care in Medicare. JAMA Internal Medicine. 2014;174(7):1067–76.

1.1 INTRODUCTION

Several recent initiatives, including the "Choosing Wisely" campaign by the American Board of Internal Medicine Foundation,[1] have focused on directly defining wasteful health care services that provide little or no health benefit to patients. It is challenging, however, to translate evidence-based lists of low-value services generated by such initiatives into meaningful metrics that can be applied to available data sources such as insurance claims.[2] The value of most services depends on the clinical situation in which they are provided, and administrative data often lack the clinical detail necessary to distinguish appropriate from inappropriate use. Consequently, the number of low-value services that can be reliably identified in claims data may be limited, and the amount of low-value care detected by claims-based measures may be highly sensitive to how the measures are defined.

Direct approaches to measuring overuse may nevertheless be useful for characterizing the potential extent of wasteful care and informing policies to address low-value practices. Indirect approaches to measuring care efficiency, such as comparing total risk-adjusted spending per patient across geographic areas or provider organizations,[3] may be challenging for policymakers and providers to act on because specific services contributing to wasteful spending are not identified.[4] Furthermore, such relative measures may fail to characterize the full extent of low-value practices if they are widespread. In contrast, direct measures could be used to identify specific instances of overuse and assess their frequency among even the most efficient providers. In addition, even a limited set of direct measures could be useful for monitoring low-value care if it reflects underlying drivers of overuse more broadly. For analogous reasons, many quality measures relating to underuse have been developed

and applied widely in quality-improvement initiatives despite similar measurement challenges.[5,6]

Drawing from evidence-based lists and the medical literature, we created algorithms to measure selected low-value services that could be applied to insurance claims data with reasonable accuracy despite the limited clinical information in claims. Using 2009 Medicare claims, we examined the use of these services and their associated spending, varying the sensitivity and specificity with which the measures likely identified overuse. We also examined whether use of different types of low-value care was correlated within regions; positive correlations might suggest that the measures reflect common drivers of overuse.

1.2 METHODS

Data Sources and Sample Population

We analyzed 2008-2009 claims data for a random 5% sample of Medicare beneficiaries, as well as demographic information from enrollment files and chronic conditions from the Chronic Condition Data Warehouse (CCW).[7] We applied measures of low-value services to 2009 claims, using 2008 claims and the CCW for relevant clinical history. Our study population consisted of 1,360,908 beneficiaries who were continuously enrolled in Part A and B of traditional fee-for-service Medicare in 2008 and while alive in 2009. We further restricted the study population to individuals who, in 2009, were living in the United States or Washington, DC, and were at least 65 years old. Our study was approved by the Harvard Medical School Committee on Human Studies and the Privacy Board of the Centers for Medicare & Medicaid Services.

Measures of Low-Value Services

We considered services that have been characterized as low-value by the American Board of Internal Medicine Foundation's Choosing Wisely initiative,[8] the US Preventive Services Task Force "D" recommendations,[9] the National Institute for Health and Care Excellence "do not do" recommendations,[10] the Canadian Agency for Drugs and Technologies in Health health technology assessments,[11] or peer-reviewed medical literature.[12] These services provide little to no clinical benefit on average, either in general or in specific clinical scenarios. From these services, we selected a subset that is relevant to the Medicare population and could be detected using Medicare claims with reasonable specificity, meaning that major clinical factors distinguishing likely overuse from appropriate use could be identified or approximated with claims and enrollment data (Appendix 1). We also required the evidence base characterizing each service as low-value to have been established before 2009. Many low-value services were not selected (e.g., imaging for pulmonary embolism without moderate or high pre-test probability[8]) because of difficulty distinguishing inappropriate from appropriate use with claims data.

For each selected service, we developed an operational definition of low-value occurrences using *Current Procedural Terminology (CPT)* codes, Berenson-Eggers Type of Service (BETOS) codes, *International Classification of Diseases, Ninth Revision (ICD-9)* diagnostic codes, CCW indicators, timing of care, site of care, and demographic information (Appendix 1). When supported by clinical evidence or guidelines, we broadened the scope of some recommendations featured in lists of low-value services. For example, we expanded the Choosing Wisely definition of low-value preoperative pulmonary testing before cardiac surgery to include pre-operative pulmonary testing

before low- or intermediate-risk surgical procedures more broadly.[13] We also combined similar low-value services (e.g. various laboratory tests for hypercoaguable states) into single measures. Table 1.1 presents the operational definitions for the 26 measures of low-value care we developed and applied to claims.

Inherent in most of our claims-based measures of low-value care was a trade-off between sensitivity (greater capture of inappropriate use) and specificity (less misclassification of appropriate use as inappropriate). To assess the variability of our findings across a spectrum of these important measurement properties, we specified two versions of each measure, one with higher sensitivity (and lower specificity) and the other with higher specificity (and lower sensitivity) for detecting low-value care (Table 1.1). Even without a gold standard for assessing service appropriateness, the relative sensitivity and specificity of our measures can be inferred from the clinical criteria we applied. For example, limiting the colorectal cancer screening measure to beneficiaries older than 85 years instead of older than 75 years decreases its sensitivity (fewer low-value instances detected) but increases its specificity (smaller proportion of appropriate services misclassified as inappropriate).

We calculated spending on low-value services using standardized prices to adjust for regional differences in Medicare payments. We used the median spending per service nationally as the standardized price for each service, including payments from Medicare, beneficiary coinsurance amounts, and any payments from other primary payers. We included related services typically bundled with the low-value service in these price estimates (e.g. contrast medium administration for an imaging study or anesthesia for a procedure). These bundles were defined based on examination of the most frequent CPT codes appearing during the day a low-value

service was provided and thus would not include subsequent care prompted by the service (e.g., further imaging for incidental findings on preoperative chest radiographs). Additional information on service detection and pricing, including the specific codes (CPT, BETOS, etc.) employed, is available in Appendix 1.

Statistical Analysis

We counted the number of times each beneficiary experienced each low-value service and calculated the per-beneficiary spending for each service. From these values, we calculated the percentage of beneficiaries receiving at least one low-value service and the aggregate spending for all beneficiaries for each service and in each of six service categories: low-value cancer screening; low-value diagnostic and preventive testing; low-value preoperative testing; low-value imaging; low-value cardiovascular testing and procedures; and other low-value surgical procedures. Aggregate spending estimates were multiplied by 20 to approximate spending for the entire Medicare population from 5% samples. We also calculated the proportion of total spending for services covered by Medicare Parts A and B (including coinsurance amounts and payments from other primary payers) devoted to services detected by low-value care measures.

We used hospital referral regions (HRRs) to examine how use of different types of low-value services was related among the same groupings of providers. Although we were not interested in geographic areas per se and although practice patterns vary within and between areas,[4] HRRs nevertheless served as a useful unit of comparison to determine whether groups of providers that were more likely to provide one type of low-value service were more likely to provide another. First, we estimated mean per-

beneficiary utilization counts in each service category at the HRR level using linear regression models with HRR fixed effects. To control for beneficiaries' sociodemographic and clinical characteristics, we included as covariates age, age squared, sex, race, indicators of 21 CCW diagnoses present before 2009 (derived from claims dating back to 1999), indicators of having multiple comorbid conditions (2 to 7+), the Rural-Urban Continuum Code for beneficiaries' county of residence, and several socioeconomic measures of the elderly population at the zip code tabulation area level (median income, percentage below the federal poverty level, and percentage with a high school diploma). To account for additional dimensions of case mix not captured by the CCW, we included indicators of conditions that qualified patients for potential receipt of several low-value services (e.g., a diagnosis of headache in 2009 qualifying beneficiaries for potentially inappropriate head imaging; see Appendix 1 for details). For each pair of low-value service categories, we then estimated correlations between regional means in adjusted use, weighted by the number of traditional fee-for-service Medicare beneficiaries in each HRR. Correlations were not substantially altered by use of random effects to estimate regional means or by the addition of indicators of qualifying conditions.

Table 1.1 Measures of Low-Value Services

Measure	Source and Supporting Literature	Operational Definition: More Sensitive, Less Specific (Base Definition)	Operational Definition: Less Sensitive, More Specific (Additional Restrictions)
Cancer Screening			
Cancer screening for patients with CKD receiving dialysis	CW[14]	Screening for cancer of the breast, cervix, colon, or prostate for patients with CKD receiving dialysis services	Only patients aged ≥75y[a]
Cervical cancer screening for women aged ≥65 y	CW, USPSTF[15]	Screening Papanicolaou test for women aged ≥65 y	No personal history of cervical cancer or dysplasia noted in claim or in prior claims[b]; no diagnoses of other female genital cancers, abnormal Papanicolaou findings, or human papillomavirus positivity in prior claims
Colorectal cancer screening for older elderly patients	USPSTF[16]	Colorectal cancer screening (colonoscopy, sigmoidoscopy, barium enema, or fecal occult blood testing) for patients aged ≥75 y	No history of colon cancer; only screening (i.e. not diagnostic) procedure codes; only patients over age 85
PSA testing for men aged ≥75 y	USPSTF[17]	PSA test for patients ≥75 y	No history of prostate cancer; only screening (i.e. not diagnostic) procedure codes
Diagnostic and Preventive Testing			
Bone mineral density testing at frequent intervals	Literature[18,19]	Bone mineral density test <2 y after prior bone mineral density test	Only patients with a diagnosis of osteoporosis prior to the initial bone mineral density test[c]
Homocysteine testing for cardiovascular disease	Literature[20]	Homocysteine testing	No diagnoses of folate or B12 deficiencies in claim and no folate or B12 testing in prior claims
Hypercoagulability testing for patients with deep vein thrombosis	CW[21]	Laboratory tests for hypercoagulable states within 30 d after diagnosis of lower-extremity deep vein thrombosis or pulmonary embolism	No evidence of recurrent thrombosis, defined by diagnosis of deep vein thrombosis or pulmonary embolism > 90 d before claim
PTH measurement for patients with stage 1-3 CKD	NICE[22,23]	PTH measurement in patients with CKD	No dialysis services before PTH testing or within 30 d after testing; no hypercalcemia diagnosis in any 2009 claim

Table 1.1 (Continued) Measures of Low-Value Services

Preoperative Testing			
Preoperative chest radiography	CADTH CW[24,25]	Chest radiograph specified as a preoperative assessment or occurring within 30 d before a low- or intermediate-risk noncardiothoracic surgical procedure[d]	No radiographs related to inpatient or emergency care[e]; only radiographs that preceded a low- or intermediate-risk noncardiothoracic surgical procedure (i.e. excluding those specified as preoperative before other procedures)[d]
Preoperative echocardiography	CW[26]	Echocardiogram specified as a preoperative assessment or occurring within 30 d before a low- or intermediate-risk noncardiothoracic surgical procedure[d]	No echocardiograms related to inpatient or emergency care[e]; only echocardiograms that preceded a low- or intermediate-risk non-cardiothoracic surgical procedure[d]
Preoperative PFT	CW[13]	PFT specified as a preoperative assessment or occurring within 30 d before a low or intermediate risk surgical procedure[f]	No PFTs related to inpatient or emergency care[e]; only PFT that preceded a low- or intermediate- risk surgical procedure[f]
Preoperative stress testing	CW[27]	Stress electrocardiography, echocardiography, or nuclear medicine imaging specified as a preoperative assessment or occurring within 30 d before a low- or intermediate-risk noncardiothoracic surgical procedure[d]	No stress testing related to inpatient or emergency care[e]; only stress testing that preceded a low- or intermediate-risk noncardiothoracic surgical procedure[d]
Imaging			
CT of the sinuses for uncomplicated acute rhinosinusitis	CW[28]	Maxillofacial CT study with a diagnosis of sinusitis in the imaging claim	No complications of sinusitis,[g] immune deficiencies, nasal polyps, or head/face trauma noted in claim; no patients with chronic sinusitis, defined by sinusitis diagnosis between 1 y and 30 d before imaging
Head imaging in the evaluation of syncope	CW NICE[29]	CT or MR imaging of the head with a diagnosis of syncope in the imaging claim	No diagnoses in claim warranting imaging[h]
Head imaging for uncomplicated headache	CW[30]	CT or MR imaging of the head with a diagnosis of (non-thunderclap, non-post-traumatic) headache	No diagnoses in claim warranting imaging[i]
EEG for headaches	CW[31]	EEG with headache diagnosis in the claim	No epilepsy or convulsions noted in current or prior claims

Table 1.1 (Continued) Measures of Low-Value Services

Back imaging for patients with nonspecific low back pain	CW, NICE[32]	Back imaging with a diagnosis of lower back pain	No diagnoses in claim warranting imaging[j]; imaging occurred within 6 wk of the first diagnosis of back pain
Screening for carotid artery disease in asymptomatic adults	CW, USPSTF[33]	Carotid imaging for patients without a history of stroke or TIA and without a diagnosis of stroke, TIA, or focal neurological symptoms in claim	Test not associated with inpatient or emergency care[k]
Screening for carotid artery disease for syncope	CW[29]	Carotid imaging with syncope diagnosis	No history of stroke or TIA; No stroke, TIA, or focal neurological symptoms noted in claim
Cardiovascular Testing and Procedures			
Stress testing for stable coronary disease	CW[34] Literature[35]	Stress testing for patients with an established diagnosis of ischemic heart disease or angina (≥6 mo before the stress test) and thus not done for screening purposes	Test not associated with inpatient or emergency care, which might be indicative of unstable angina [k]; only patients with a past diagnosis of myocardial infarction in order to exclude patients with a history of noncardiac chest pain inaccurately coded as angina (i.e., those with no underlying ischemic heart disease who might benefit from screening and optimization of medical management)
Percutaneous coronary intervention with balloon angioplasty or stent placement for stable coronary disease	Literature[35,36]	Coronary stent placement or balloon angioplasty for patients with an established diagnosis of ischemic heart disease or angina (≥6 mo before the procedure);procedure not associated with an ED visit,[k] which might be indicative of acute coronary syndrome	Only patients with a past diagnosis of myocardial infarction in order to exclude patients with a history of non-cardiac chest pain inaccurately coded as angina
Renal artery angioplasty or stenting	Literature[37,38]	Renal/visceral angioplasty or stent placement	Diagnosis of renal atherosclerosis or renovascular hypertension noted in procedure claim
Carotid endarterectomy in asymptomatic patients	CW[33,39]	Carotid endarterectomy for patients without a history of stroke or TIA and without stroke, TIA, or focal neurological symptoms noted in claim	Operation not associated with an ED visit[k]; only female patients[l]
IVC filters for the prevention of pulmonary embolism	Literature[40,41]	Any IVC filter placement	No additional restrictions

Table 1.1 (Continued) Measures of Low-Value Services

Other Surgery			
Vertebroplasty or kyphoplasty for osteoporotic vertebral fractures	Literature[42-45]	Vertebroplasty/kyphoplasty for vertebral fracture	No bone cancers, myeloma, or hemangioma noted in procedure claim
Arthroscopic surgery for knee osteoarthritis	NICE[46,47]	Arthroscopic debridement/chondroplasty of the knee	Diagnosis of osteoarthritis or chondromalacia in the procedure claim; no meniscal tear noted in the procedure claim

Abbreviations: CADTH, Canadian Agency for Drugs and Technologies in Health health technology assessments; CKD, chronic kidney disease; CT, computed tomography; CW, Choosing Wisely; ED, emergency department; EEG, electroencephalography; IVC, inferior vena cava; MR, magnetic resonance; NICE, National Institute for Health and Care Excellence "do not do" list; PFT, pulmonary function testing; PSA, prostate-specific antigen; PTH, parathyroid hormone; TIA, transient ischemic attack; USPSTF, US Preventive Services Task Force C or D recommendations.

[a] This age cutoff is included because the distribution of kidney transplant recipient ages within the sample suggests transplantation is uncommon in patients 75 years or older.

[b] Throughout the table, "prior claims" refers to all claims from January 1, 2008, until 1 d before the service of interest.

[c] This restriction limits the measure to testing of patients with osteoporosis.

[d] Including breast procedures, colectomy, cholecystectomy, transurethral resection of the prostate, hysterectomy, orthopedic surgical procedures other than hip and knee replacement, corneal transplant, cataract removal, retinal detachment, hernia repair, lithotripsy, arthroscopy, and cholecystectomy. The 30-day window between preoperative testing and surgery was derived empirically based on distribution of intervals between test and procedure.

[e] Inpatient-associated is defined here as occurring during within 30 d after an inpatient stay; ED-associated, during or 1 d after an ED visit.

[f] Including procedures listed in footnote d as well as coronary artery bypass graft, aneurysm repair, thromboendarterectomy, percutaneous transluminal
coronary angioplasty, and pacemaker insertion.

[g] Complications of sinusitis include eyelid inflammation, acute inflammation of orbit, orbital cellulitis, and visual problems.

[h] Exclusion diagnoses include epilepsy, giant cell arteritis, head trauma, convulsions, altered mental status, nervous system symptoms (eg, hemiplegia), disturbances of skin sensation, speech problems, stroke, transient ischemic attack, and history of stroke.

Table 1.1 (Continued) Measures of Low-Value Services

[i] Exclusion diagnoses include those listed in the preceding footnote as well as cancer and history of cancer.

[j] Exclusion diagnoses include cancer, trauma, intravenous drug abuse, neurological impairment, endocarditis, septicemia, tuberculosis, osteomyelitis, fever, weight loss, loss of appetite, night sweats, and anemia.

[k] Inpatient-associated is defined here as occurring during an inpatient stay; ED-associated, during or within 14 d after an ED visit.

[l] Restriction is based on sex-specific subgroup analyses of procedure efficacy in the referenced literature.

1.3 RESULTS

Among 1,360,908 beneficiaries in the study sample, 1,094,374 instances of care provision (80 services per 100 beneficiaries) were detected by the more sensitive measures of low-value services, corresponding to 21.9 million instances for the entire traditional Medicare population in 2009. Forty-two percent of beneficiaries received at least 1 service detected by the more sensitive measures. Our more specific but less sensitive measures of low-value care detected 454,783services (33 per 100 beneficiaries), corresponding to 9.1 million services for the entire Medicare population. Twenty-five percent of beneficiaries received at least 1 of these services.

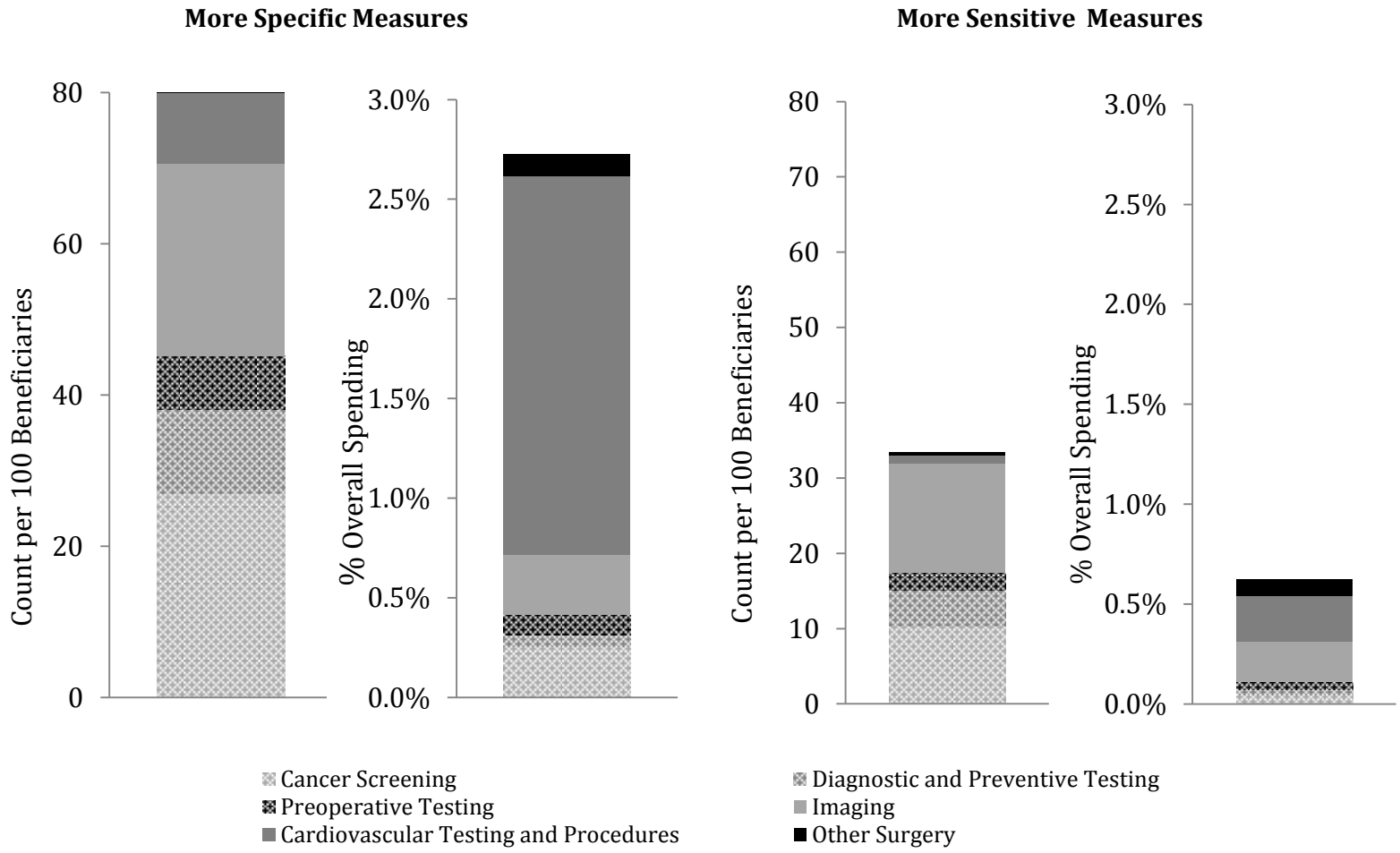
Spending for services detected by our more sensitive measures of low-value care totaled $8.5 billion for the entire Medicare population, or $310 per beneficiary, while spending for services detected by our more specific measures totaled $1.9 billion, or $71 per beneficiary. These amounts comprised 2.7% and 0.6%, respectively, of total annual spending in 2009 on services covered by Part A and B of Medicare.

Figure 1.1 presents utilization rates and their associated spending, decomposed by category of low-value care measures. Imaging, cancer screening, and diagnostic and preventive testing measures detected most of the use, whereas measures of imaging and cardiovascular testing and procedures detected most of the spending (see Appendix 1 for these results in tabular form). Table 1.2 presents utilization rates and associated spending captured by each of the 26 measures of low-value care. Individual measures with major contributions to spending included both high-price, low-utilization items such as percutaneous coronary intervention for stable coronary disease and low-price, high-utilization items such as screening for asymptomatic carotid artery disease.

Table 1.3 presents correlations between adjusted levels of regional service use in different categories of low-value care, as detected by our more sensitive measures. Per-beneficiary utilization counts were positively correlated with one another for five of the six categories. Correlation coefficients ranged from 0.14 to 0.54 across all pair-wise combinations of these five categories ($P \leq 0.01$), with a mean of 0.33. Non-cardiovascular surgical procedures were not positively correlated with use in other categories of measures. The measures exhibited good internal consistency across all categories (Chronbach's alpha, 0.68).

Adjusted regional spending on services detected by more sensitive measures of low-value care ranged from $227 per-beneficiary in the 5th percentile to $416 per-beneficiary in the 95th percentile of HRRs (median, $304; inter-quartile range, $272 to $343). Thus, low-value spending detected in regions at the 5th percentile of the regional distribution exceeded the difference in detected low-value spending between regions at the 5th and 95th percentiles ($189/beneficiary).

Figure 1.1 Utilization Rates and Associated Spending for Services Detected by Low-Value Care Measures Among Medicare Beneficiaries in 2009



Count refers to unique incidences of service provision; overall spending, total spending on all services covered by Medicare Parts A and B (see Table 1.1 for services included in each category and for operational definitions of all measures).

Table 1.2 Service Counts and Associated Spending Detected by Measures of Low-Value Care

	More Sensitive Versions of Measures						More Specific Version of Measures					
Measure (Abbreviated)	Count per 100 benes[a]	% of low-value count	% of benes affected	Spending ($M)	% of low-value spending	% of overall spending[b]	Count per 100 benes[a]	% of low-value count	% of benes affected	Spending ($M)	% of low-value spending	% of overall spending[b]
Imaging for non-specific low back pain	12.4	15%	9.4%	226	3%	0.07%	4.5	14%	4.1%	82	4%	0.03%
PSA screening at age >75 y	12.0	15%	8.3%	98	1%	0.03%	2.8	8%	2.7%	23	1%	0.01%
PTH testing in early CKD	7.9	10%	2.5%	137	2%	0.04%	3.1	9%	1.7%	53	3%	0.02%
Stress testing for stable coronary disease	7.8	10%	7.3%	2,065	24%	0.67%	0.8	2%	0.8%	212	11%	0.07%
Colon cancer screening for older elderly patients	7.7	10%	6.9%	573	7%	0.18%	0.9	3%	0.8%	7	0%	0.00%
Cervical cancer screening at age > 65 y	7.0	9%	6.9%	120	1%	0.04%	6.5	19%	6.4%	111	6%	0.04%
Carotid artery disease screening for asymptomatic patients	6.6	8%	6.0%	323	4%	0.10%	5.6	17%	5.1%	274	14%	0.09%
Preoperative radiography	5.5	7%	5.1%	75	1%	0.02%	1.6	5%	1.6%	22	1%	0.01%
Head imaging for headache	3.4	4%	3.1%	211	2%	0.07%	2.4	7%	2.2%	146	8%	0.05%
Homocysteine testing for cardiovascular disease	2.0	3%	1.5%	15	0%	0.00%	0.8	2%	0.6%	6	0%	0.00%
Head imaging for syncope	1.4	2%	1.3%	85	1%	0.03%	1.0	3%	0.9%	60	3%	0.02%
Bone mineral density testing at frequent intervals	1.0	1%	1.0%	20	0%	0.01%	0.8	3%	0.8%	17	1%	0.01%
Carotid artery disease screening for syncope	1.0	1%	1.0%	49	1%	0.02%	0.7	2%	0.7%	33	2%	0.01%
PCI/stenting for stable coronary disease	0.8	1%	0.7%	2,810	33%	0.91%	0.1	0%	0.1%	212	11%	0.07%

Table 1.2 (Continued) Service Counts and Associated Spending Detected by Measures of Low-Value Care

Preoperative echocardiography	0.8	1%	0.8%	58	1%	0.02%	0.3	1%	0.3%	21	1%	0.01%
Preoperative stress testing	0.7	1%	0.7%	180	2%	0.06%	0.3	1%	0.3%	81	4%	0.03%
CT scan for rhinosinusitis	0.6	1%	0.6%	42	1%	0.01%	0.3	1%	0.3%	23	1%	0.01%
Renal artery stenting	0.4	0%	0.3%	705	8%	0.23%	0.1	0%	0.1%	139	7%	0.04%
Vertebroplasty	0.3	0%	0.3%	199	2%	0.06%	0.3	1%	0.3%	196	10%	0.06%
Arthroscopic surgery for knee osteoarthritis	0.2	0%	0.2%	143	2%	0.05%	0.1	0%	0.1%	63	3%	0.02%
Cancer screening for patients with CKD receiving dialysis	0.2	0%	0.2%	4	0%	0.00%	0.1	0%	0.1%	1	0%	0.00%
IVC filter placement	0.2	0%	0.2%	43	1%	0.01%	0.2	1%	0.2%	43	2%	0.01%
Preoperative PFT	0.2	0%	0.2%	2	0%	0.00%	0.1	0%	0.1%	1	0%	0.00%
Carotid endarterectomy for asymptomatic patients	0.1	0%	0.1%	263	3%	0.08%	0.1	0%	0.0%	110	6%	0.04%
Hypercoagulability testing after DVT	0.1	0%	0.1%	3	0%	0.00%	0.0	0%	0.0%	1	0%	0.00%
EEG for headache	0.1	0%	0.1%	3	0%	0.00%	0.0	0%	0.0%	2	0%	0.00%
Total	80.4	100%	42%[c]	8,451	100%	2.7%	33.4	100%	25%[c]	1,941	100%	0.6%

Abbreviations: Bene, Beneficiaries; CKD, chronic kidney disease; CT, computed tomography; DVT, deep vein thrombosis; EEG, electroencephalography; IVC, inferior vena cava; PBA, proportion of beneficiaries affected; PCI, percutaneous coronary intervention; PFT, pulmonary function testing; PLVC, proportion of low-value count; PLVS, proportion of low-value spending; POS, proportion of overall spending; PSA, prostate-specific antigen; PTH, parathyroid hormone.

[a] Count refers to the number of unique incidences of service provision.
[b] Overall spending refers to annual spending for services covered by Medicare Parts A and B. See Table 1 for service category assignments and for operational definitions of all measures.
[c] Totals do not equal column sums because some patients received multiple services.

Table 1.3 Correlations in Regional Use Between Categories of Measures of Low-Value Care

Category	Cancer Screening	Diagnostic and Preventive Testing	Preoperative Testing	Imaging	Cardiovascular Testing and Procedures	Other Surgery
Cancer Screening	1 [Reference]					
Diagnostic and Preventive Testing	0.35[b]	1 [Reference]				
Preoperative Testing	0.32[b]	0.14[c]	1 [Reference]			
Imaging	0.50[b]	0.32[b]	0.31[b]	1 [Reference]		
Cardiovascular Testing and Procedures	0.29[b]	0.29[b]	0.27[b]	0.54[b]	1 [Reference]	
Other Surgery	−0.14[c]	-0.07	-0.16[b]	0.01	0.06	1 [Reference]

[a] Values represent Pearson correlation coefficients
[b] P<.01
[c] P<.05

1.4 Discussion

In this national study of selected low-value services, Medicare beneficiaries commonly received care that was likely to provide minimal or no benefit on average. Even when applying narrower versions of our limited number of measures of overuse, we identified low-value care affecting one-quarter of Medicare beneficiaries. These findings are consistent with the notion that wasteful practices are pervasive in the US health care system.

Within regions, different types of low-value use generally exhibited significantly positive correlations with one another, ranging from weak to moderate in strength, although one category of low-value use (non-cardiovascular surgical procedures) was not positively correlated with the others. These findings suggest that many low-value services may be driven by common factors. Therefore, claims-based measures, although limited in number and the amount of wasteful spending they detect, could be useful for monitoring low-value care more broadly, including some care that may be difficult to measure with claims.

Although these findings suggest that direct approaches to measuring wasteful care may be tractable and informative, other findings underscore potential challenges in developing and applying direct measures of overuse. In particular, the amount of low-value care we detected varied substantially with the clinical specificity of our measures. Estimates of the proportion of Medicare beneficiaries receiving at least one measured low-value service decreased from 42% to 25% when we used more restrictive definitions that traded off sensitivity for specificity, and the contribution of low-value spending to total spending decreased from 2.7% to 0.6%. For example, our more sensitive measure of low-value imaging for low back pain captured more inappropriate

use of imaging studies at the expense of including some appropriate use. Our more specific measure was less likely to include appropriate use but probably excluded many low-value studies, as suggested by the 3-fold reduction in the number of studies captured.

Thus, the performance of administrative rules to reduce overuse through coverage policy, cost-sharing, or value-based payment (e.g., pay for performance) may depend heavily on measure definition. Such strategies may be appropriate for select services whose value is invariably low or whose low-value applications can be identified with high reliability. For other services, however, more sensitive measures could result in unintended restriction of appropriate tests and procedures by coverage and payment policies, whereas more specific measures could substantially limit the effect of these strategies. Provider groups seeking to minimize wasteful spending— for example, in response to global budgets— may be able to distinguish appropriate from inappropriate practices at the point of care without having to use rigid rules derived from incomplete clinical data.

We also found that, although spending on low-value services varied considerably across regions, spending on low-value services was substantial even in regions where it was lowest. For example, low-value spending at the 5th percentile of the regional distribution of low-value spending was greater than the difference in low-value spending between the 5th and 95th percentiles. This finding suggests potential advantages of direct measurement over relative spending comparisons as a basis for detecting overuse because overuse may be substantial even among more efficient providers.

Our study has several limitations. Most notably, we analyzed only 26 measures of low-value services. In selecting these measures, we emphasized the specificity with which overuse could be detected with claims data and created more restrictive versions that limited contributions of potentially valuable service use to low-value spending totals and utilization counts. Despite the limited number of services we examined, their frequency and correlations with one another suggest substantial and widespread wasteful care. Use of a broader set of less specific and more sensitive measures would capture more low-value care. Similarly, broader definitions of wasteful spending that include downstream costs of low-value service use (e.g., repeat imaging for incidental findings) would capture more spending than our measures did. For example, one study estimated that testing costs may account for just 2% of the lifetime costs of prostate-specific antigen screening.[48]

Clinical data from linked medical records might support a more extensive assessment of the properties of claims-based measures. However, we would not expect the incorporation of more detailed data to substantially alter the amount of low-value care captured by many of our measures (e.g. cancer screening in patients above certain ages, inappropriately frequent bone mineral density testing, homocysteine testing for cardiovascular disease, renal artery stenting, and vertebroplasty). Furthermore, by varying the definitions of our measures, we were able to demonstrate potential limitations of claims-based measures without having to use medical record data; any inconsistencies between claims and medical records in the amount of low-value care detected would have similar implications for strategies to address wasteful practices. Moreover, we focused on the potential utility of claims-based measures because medical record review as a means to measure and monitor wasteful care is costly and

thus not feasible on a large scale. Nevertheless, validation of claims-based measures against a gold standard of clinical appropriateness will be needed to more precisely define their strengths and weaknesses and assess their utility for different purposes, such as monitoring, profiling, payment policy, or coverage design.

Although our analysis suggests that common drivers of low-value care exist, our study did not identify specific determinants of wasteful care. Factors associated with low-value care may also be associated with high-value care.[49,50] Coupling measures of overuse with measures of underuse may therefore be important when evaluating programs intended to achieve more cost-effective care.

Finally, unmeasured variation in diagnostic coding practices or case mix may have contributed to positive correlations between regional use of different low-value services in our study. These were not likely sources of significant bias, however, because we found a significant positive correlation between categories of low-value services that did not rely on diagnosis codes to define (i.e. age-inappropriate cancer screening and preoperative testing) and because our results were not sensitive to adjustment for additional conditions qualifying beneficiaries for potential receipt of several low-value services.

Many quality measures have been developed to assess underuse but few to assess overuse. Our study findings illustrate the potential utility and limitations of a direct approach to detect wasteful care. Despite their imperfections, claims-based measures of low-value care could be useful for tracking overuse and evaluating programs to reduce it. However, many direct claims-based measures of overuse may be insufficiently accurate to support targeted coverage or payment policies that have a meaningful effect on use without resulting in unintended consequences. Broader

payment reforms such as global or bundled payment models could allow greater provider discretion in defining and identifying low-value services while incentivizing their elimination.

1.5 References

1. Cassel CK, Guest JA. Choosing Wisely: helping physicians and patients make smart decisions about their care. *JAMA*. 2012; 307(17):1801-1802.
2. Elshaug AG, McWilliams JM, Landon BE. The value of low-value lists. *JAMA*. 2013; 309(8):775-776.
3. Skinner JS. Causes and consequences of regional variations in health care. In: Pauly MV, McGuire T, Barros PP, eds. *Handbook of Health Economics*. Amsterdam, the Netherlands: North-Holland/ Elsevier; 2012:45-94.
4. Newhouse JP, Garber AM, Graham RP, McCoy MA, Mancher M, Kibria A. *Variation in Health Care Spending: Target Decision Making, Not Geography.* Washington, DC: Institute of Medicine; 2013.
5. Wachter RM. Expected and unanticipated consequences of the quality and information technology revolutions. *JAMA*. 2006; 295(23): 2780-2783.
6. National Committee for Quality Assurance. The state of health care quality 2012: focus on obesity and on Medicare plan improvement. http://www.ncqa.org/Portals/0/State%20of%20Health%20Care/2012/SOHC_Report_Web.pdf. Accessed April 4, 2014.
7. Center for Medicare & Medicaid Services. Chronic Conditions Data Warehouse. http://www.ccwdata.org/. Published 2013. Accessed April 4, 2014.
8. American Board of Internal Medicine Foundation. Choosing Wisely: lists of five things physicians and patients should question. http://www.choosingwisely.org/doctor-patient-lists/. Published 2013. Accessed April 4, 2014.

9. U.S. Preventive Services Task Force. Recommendations for adults. http://www.uspreventiveservicestaskforce.org/adultrec.htm. Published 2013. Accessed April 4, 2014.
10. National Institute for Health and Care Excellence. NICE "do not do" recommendations. http://www.nice.org.uk/usingguidance/donotdorecommendations/index.jsp. Published 2011. Accessed April 4, 2014.
11. Canadian Agency for Drugs and Technologies in Health. Health technology assessments. http://cadth.ca/en/products/health-technology-assessment. Accessed April 4, 2014.
12. Elshaug AG, Watt AM, Mundy L, Willis CD. Over 150 potentially low-value health care practices: an Australian study. *Med J Aust.* 2012; 197(10):556-560.
13. Qaseem A, Snow V, Fitterman N, et al; Clinical Efficacy Assessment Subcommittee of the American College of Physicians. Risk assessment for and strategies to reduce perioperative pulmonary complications for patients undergoing noncardiothoracic surgery: a guideline from the American College of Physicians. *Ann Intern Med.* 2006; 144(8):575-580.
14. Holley JL. Screening, diagnosis, and treatment of cancer in long-term dialysis patients. *Clin J Am Soc Nephrol.* 2007; 2(3):604-610.
15. Vesco K, Whitlock E, Eder M, et al. *Screening for Cervical Cancer: A Systematic Evidence Review for the U.S. Preventive Services Task Force*. Rockville, MD: Agency for Healthcare Research and Quality; 2011.
16. Whitlock EP, Lin JS, Liles E, Beil TL, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. *Ann Intern Med.* 2008; 149(9):638-658.

17. Lin K, Lipsitz R, Miller T, Janakiraman S; U.S. Preventive Services Task Force. Benefits and harms of prostate-specific antigen screening for prostate cancer: an evidence update for the U.S. Preventive Services Task Force. *Ann Intern Med.* 2008; 149(3):192-199.
18. Bell KJL, Hayen A, Macaskill P, et al. Value of routine monitoring of bone mineral density after starting bisphosphonate treatment: secondary analysis of trial data. *BMJ*. 2009; 338:b2266.
19. Hillier TA, Stone KL, Bauer DC, et al. Evaluating the value of repeat bone mineral density measurement and prediction of fractures in older women: the study of osteoporotic fractures. *Arch Intern Med.* 2007; 167(2):155-160.
20. Martí-Carvajal AJ, Solà I, Lathyris D, Karakitsiou D-E, Simancas-Racines D. Homocysteine-lowering interventions for preventing cardiovascular events. *Cochrane Database Syst Rev.* 2013;1:CD006612.
21. Baglin T, Gray E, Greaves M, et al; British Committee for Standards in Haematology. Clinical guidelines for testing for heritable thrombophilia. *Br J Haematol.* 2010; 149(2):209-220.
22. Levin A, Bakris GL, Molitch M, et al. Prevalence of abnormal serum vitamin D, PTH, calcium, and phosphorus in patients with chronic kidney disease: results of the study to evaluate early kidney disease. *Kidney Int.* 2007; 71(1):31-38.
23. Palmer SC, McGregor DO, Craig JC, Elder G, Macaskill P, Strippoli GF. Vitamin D compounds for people with chronic kidney disease not requiring dialysis. *Cochrane Database Syst Rev.* 2009; (4):CD005633.

24. Mohammed T, Kirsch J, Amorosa J, et al. *ACR Appropriateness Criteria Routine Admission and Preoperative Chest Radiography.* Reston, VA: American College of Radiology; 2011.
25. Joo HS, Wong J, Naik VN, Savoldelli GL. The value of screening preoperative chest x-rays: a systematic review*. Can J Anaesth.* 2005; 52(6):568-574.
26. Douglas PS, Garcia MJ, Haines DE, et al; ACCF/ASE/AHA/ASNC/HFSA/HRS/ SCAI/SCCM/SCCT/SCMR 2011 Appropriate use criteria for echocardiography. *J Am Coll Cardiol.* 2011; 57(9):1126-1166.
27. Fleisher LA, Beckman JA, Brown KA, et al;. ACC/AHA 2007 guidelines on perioperative cardiovascular evaluation and care for noncardiac surgery. *Circulation*. 2007; 116(17):e418-e499.
28. Cornelius RS, Martin J, Wippold FJ II, et al; ACR Appropriateness Criteria sinonasal disease*. J Am Coll Radiol*. 2013; 10(4):241-246.
29. Moya A, Sutton R, Ammirati F, et al; Guidelines for the diagnosis and management of syncope (version 2009). *Eur Heart J.* 2009; 30(21):2631-2671.
30. Jordan J, Wippold FI, Cornelius R, et al. *ACR Appropriateness Criteria Headache*. Reston, VA: American College of Radiology; 2009.
31. Gronseth GS, Greenberg MK. The utility of the electroencephalogram in the evaluation of patients presenting with headache: a review of the literature. *Neurology*. 1995; 45(7):1263-1267.
32. Chou R, Fu R, Carrino JA, Deyo RA. Imaging strategies for low-back pain: systematic review and meta-analysis. *Lancet*. 2009; 373(9662):463-472.

33. Wolff T, Guirguis-Blake J, Miller T, Gillespie M, Harris R. Screening for carotid artery stenosis: an update of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med.* 2007; 147(12):860-870.
34. Hendel RC, Berman DS, Di Carli MF, et al; ACCF/ASNC/ACR/AHA/ASE/SCCT/ SCMR/SNM 2009 appropriate use criteria for cardiac radionuclide imaging. *Circulation*. 2009; 119(22):e561-e587.
35. Boden WE, O'Rourke RA, Teo KK, et al; COURAGE Trial Research Group. Optimal medical therapy with or without PCI for stable coronary disease. *N Engl J Med.* 2007; 356(15):1503-1516.
36. Lin GA, Dudley RA, Redberg RF. Cardiologists' use of percutaneous coronary interventions for stable coronary artery disease. *Arch Intern Med.* 2007; 167(15):1604-1609.
37. Wheatley K, Ives N, Gray R, et al; ASTRAL Investigators. Revascularization versus medical therapy for renal-artery stenosis. *N Engl J Med*. 2009; 361(20):1953-1962.
38. Cooper CJ, Murphy TP, Cutlip DE, et al; CORAL Investigators. Stenting and medical therapy for atherosclerotic renal-artery stenosis. *N Engl J Med*. 2014; 370(1):13-22.
39. Goldstein LB, Bushnell CD, Adams RJ, et al; Guidelines for the primary prevention of stroke. *Stroke*. 2011; 42(2):517-584.
40. PREPIC Study Group. Eight-year follow-up of patients with permanent vena cava filters in the prevention of pulmonary embolism: the PREPIC (Prevention du Risque d'Embolie Pulmonaire par Interruption Cave) randomized study. *Circulation*. 2005; 112(3):416-422.

41. Sarosiek S, Crowther M, Sloan JM. Indications, complications, and management of inferior vena cava filters: the experience in 952 patients at an academic hospital with a level I trauma center. *JAMA Intern Med*. 2013; 173(7):513-517.
42. Kallmes DF, Comstock BA, Heagerty PJ, et al. A randomized trial of vertebroplasty for osteoporotic spinal fractures. *N Engl J Med*. 2009; 361(6):569-579.
43. Buchbinder R, Osborne RH, Ebeling PR, et al. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. *N Engl J Med.* 2009; 361(6):557-568.
44. Boonen S, Van Meirhaeghe J, Bastian L, et al. Balloon kyphoplasty for the treatment of acute vertebral compression fractures: 2-year results from a randomized trial. *J Bone Miner Res.* 2011; 26(7):1627-1637.
45. McCullough BJ, Comstock BA, Deyo RA, Kreuter W, Jarvik JG. Major medical outcomes with spinal augmentation vs conservative therapy. *JAMA Intern Med.* 2013; 173(16):1514-1521.
46. Laupattarakasem W, Laopaiboon M, Laupattarakasem P, Sumananont C. Arthroscopic debridement for knee osteoarthritis. *Cochrane Database Syst Rev.* 2008; (1):CD005118. doi:10.1002/14651858.CD005118.pub2.
47. Katz JN, Brophy RH, Chaisson CE, et al. Surgery versus physical therapy for a meniscal tear and osteoarthritis. *N Engl J Med*. 2013; 368(18):1675-1684.
48. Shteynshlyuger A, Andriole GL. Cost-effectiveness of prostate specific antigen screening in the United States: extrapolating from the European Study of Screening for Prostate Cancer. *J Urol.* 2011; 185(3):828-832.
49. Landrum MB, Meara ER, Chandra A, Guadagnoli E, Keating NL. Is spending more always wasteful? the appropriateness of care and outcomes among colorectal cancer patients. *Health Aff (Millwood).* 2008; 27(1):159-168.

50. Leape LL, Park RE, Solomon DH, Chassin MR, Kosecoff J, Brook RH. Does inappropriate use explain small-area variations in the use of health care services? *JAMA*. 1990; 263(5):669-672.

Chapter 2

Low-Value Care in Provider Organizations

2.1 Low-Value Care in Provider Organizations: Variation, Persistence, and Consistency

2.1.1 Introduction

Provider organizations are an increasing focus of initiatives that aim to reduce unnecessary health care utilization. For example, both private and public insurers have pursued accountable care organization (ACO) programs, which base payments to provider organizations on total patient spending relative to a global budget.[1,2] How organization-level incentives will affect patient care depends on how these organizations influence physician behavior.[3] From small group practices to large integrated delivery systems, provider organizations may shape the practice patterns of affiliated physicians in several ways: by setting the form of physician compensation,[4] by investing in care inputs like clinical decision support[5,6] or in delivery models like the patient-centered medical home,[7] by fostering social networks of peer physicians,[8] or by selectively recruiting physicians based on their training background.[9–11] Alternatively, loose organizational ties, which may be created to improve provider market share, may not meaningfully affect patient care.

We explore whether patterns of low-value service use are consistent with provider organizations influencing the value of care that patients receive. Specifically, we examined whether provider organizations exhibited a profile of overuse that is measurable based on administrative claims data, like the spending or quality profiles investigated for physicians within the same region or hospital.[12–14] Because it is difficult to distinguish between high-value and low-value services in many clinical scenarios, our methods drew from recent efforts by specialty societies to identify services that provide

minimal patient benefit.[15] Using 2007-2011 Medicare fee-for-service claims data, we study 31 of these services. Specifically, we measure three properties of low-value service use in provider organizations: variation across organizations, persistence of service use within an organization over time, and the consistency of organizational behavior across different types of low-value services.

2.1.2 METHODS

Study Population of Patients and Organizations

Our primary data were 2007-2011 claims and enrollment information for a 20% random annual sample of Medicare fee-for-service beneficiaries. In each year of the study period, beneficiaries were excluded from the sample if they were not continuously enrolled in Medicare Parts A and B (while alive) during that year and during the prior year. The prior year of enrollment was necessary because the detection of certain low-value services depends on diagnoses and procedures found in prior claims. Beneficiaries were also excluded from the study sample for any years in which they did not receive primary care services, which were necessary for attribution to a provider organization.

We constructed two different samples in order to characterize practice patterns for large provider organizations in general and for organizations joining ACO programs in particular. For the first sample, which we refer to as the general sample, provider organizations were defined by a single taxpayer identification number (TIN). TINs, which are included in Medicare claims for professional services, can be shared by multiple physicians and typically identify group practices or broader provider organizations.[16] We restricted the general sample to larger organizations, specifically

those organizations to which we attributed 1,000 or more patient-years during the study period.

For the second sample, the ACO sample, an organization was defined as the collection of TINs for providers that formed a Medicare ACO in 2012 or 2013. The ACO sample included 32 organizations that participated in the Medicare Pioneer ACO Program in 2012 and 218 organizations from the Medicare Shared Savings Program (MSSP), 114 that entered in 2012 and 104 that entered in 2013. To identify each ACO, we matched publicly available lists of ACOs' participating practices and facilities to TINs using public databases.[17,18] Lists of ACO participants also contained individual affiliated physicians, which we matched to the most common TIN included in each physician's 2011-2012 Medicare claims. Using TIN-based definitions of ACOs allowed for a consistent organizational definition over the five-year study period despite turnover of physicians within an ACO. We did not measure low-value service use in 2012, when ACO contracts began, so that our results would not reflect practice pattern changes associated with the contracts.

Following previously described methods,[19] each beneficiary was attributed to an organization based on MSSP rules for patient attribution. Beneficiaries were attributed to organizations that accounted for the most allowed charges for outpatient primary care services during the year. Attribution was performed separately for the general sample and for the ACO sample. For ACO attribution, beneficiaries were not attributed to an ACO if they accumulated more primary care charges at a non-ACO TIN than at an ACO.

Low-Value Services

Composite measures of each organization's use of low-value services were constructed based on 31 low-value services.[20] As described in a previous study of 26 of these services,[20] these services were chosen because they provide minimal average clinical benefit in specific clinical scenarios. The services were selected from evidence-based lists published in the American Board of Internal Medicine Foundation's Choosing Wisely initiative,[21] the US Preventive Services Task Force "D" recommendations,[22] and the Canadian Agency for Drugs and Technologies in Health technology assessments,[23] or from the peer-reviewed medical literature.[24] Services were excluded from measure construction if their appropriate use could not be distinguished from likely inappropriate overuse with reasonable accuracy using Medicare claims and enrollment data. Because there can be scope for discretion in how to define a low-value service,[20] we tended to employ more specific definitions of low-value services that reduce the likelihood of classifying a high-value service as low-value.

Table 2.1.1 presents the operational definitions used to detect each type of low-value service. These definitions incorporate relevant patient demographic or clinical characteristics like age, sex, and current or past diagnoses. Some measure definitions also rely on the timing of a service (e.g. imaging preceding a surgical operation) or the service setting (e.g. non-emergent). Service occurrences meeting these definitions were detected on the basis of information in claims like Current Procedural Terminology (CPT) service procedure codes and International Classification of Diseases, Ninth Revision (ICD-9) patient diagnosis codes, as well as information in the annual enrollment file, like age and presence of chronic conditions. We employed claims data from as early as January 1 of the year before a service occurred in order to evaluate

whether the service met the measure definition. Details regarding service detection, including all codes used in detection algorithms, are presented in the supporting materials (Appendix 2).

Because some low-value services do not apply to all beneficiaries, we defined denominator criteria for each service. For example, to qualify for the denominator for preoperative testing services, beneficiaries must have undergone surgery (Table 2.1.1). These denominator criteria were used to adjust organizations' measured rates of a low-value service for the number of beneficiaries within that organization who could possibly receive the service. We attempted to avoid denominator criteria that might be sensitive to variation in organizations' diagnostic coding practices. For example, the denominator for the detection of head imaging for an uncomplicated headache was not restricted to patients with diagnoses of uncomplicated headache, since that diagnosis may be coded with varying completeness across organizations.

Covariates

We adjusted organizations' rates of low-value service delivery for several patient characteristics in the annual Medicare enrollment file: age, age-squared, race/ethnicity, sex, hospital referral region (HRR), disability as the initial reason for Medicare entitlement, diagnosis of end-stage renal disease, and diagnosis of chronic conditions recorded in the Chronic Condition Warehouse (CCW). CCW conditions are drawn from diagnoses in Medicare claims from as early as 1999. We created binary indicators for the presence of each condition prior to the study year, and indicators for the total count of conditions, top-coded at nine. HRR was determined based on beneficiary ZIP code.[25] We also obtained the following characteristics of the population

aged 65 and over in each beneficiary's local area: median income, fraction of residents below the federal poverty level (FPL), fraction of residents with a high school degree, and fraction of residents with a college degree. These characteristics, all measured at the level of the ZIP code tabulation area (ZCTA), were obtained from the 2007-2011 American Community Survey Summary File.

Statistical Analysis

We estimated three characteristics of organizations' patterns of low-value service use: (1) variation across organizations in the total number of low-value services, (2) the persistence of organizations' levels of low-value services over time, and (3) the correlation between organizations' use of different categories of low-value services. Constructing each estimate involved three general steps. First, we adjusted organizations' use of each of the 31 low-value services for case mix. Second, these adjusted scores for each service were combined to create composite scores of each organization's overall low-value service use. Third, we produced the parameters of interest by fitting random effects models to the composite scores. This approach follows established practices of analyzing composite measures of quality that are based on multiple quality components.[26] Details regarding these methods, briefly described below, are presented in the supporting materials.

Organizations' use of each low-value service was adjusted for case mix using ordinary least squares models of the following form:

$$Y_{ijkt} = \beta_0 + \beta_1 Covariates_{it} + \beta_2 Year\_indicators_t + \epsilon_{ijkt}$$

with Y_{ijkt} denoting the count of low-value service k during year t for beneficiary i, who was assigned to organization j. *Covariates* is a vector of the beneficiary characteristics

listed above, including indicators for beneficiary HRR, and *Year_indicators* are indicators for each year. Every organization's case mix-adjusted score was calculated for each service based on the error terms ϵ_{ijtk} for attributed beneficiaries. Note that estimating a separate model for each low-value service allowed for service-specific case mix adjustment. Each model included data from only those beneficiaries who satisfied the denominator condition for the service. When analyzing the ACO sample, we included observations from those beneficiaries who accumulated more primary care charges at a non-ACO TIN than an ACO, preventing their attribution to an ACO. Including these additional beneficiaries allowed us to adjust for regional factors even in regions served by only a single ACO. In order to explore the amount of organizational variation that could be accounted for by regional factors, we repeated the above regressions without including HRR indicators as covariates.

Organizations' composite measures of low-value use were calculated as a weighted sum of risk-adjusted scores for multiple services. The weighting method ensured that, for every service, an increase in the risk-adjusted count of that service would contributed to an equal increase in the organization's composite measure. When estimating variation in the overall use of low-value services, we constructed a single composite measure that encompassed an organization's use of all low-value services. When estimating persistence in organizational behavior, we constructed one composite measure for 2010 and one for 2011 (supporting materials). When estimating organizational consistency, we constructed composite measures for each of the six clinical categories of low-value services in order to estimate correlations between these measures (Table 2.1.1).

We used random effects modeling to estimate variation, persistence, and consistency parameters (see Appendix 2). We chose random effects model because they produce parameter estimates that account for sampling error stemming from finite sample sizes.[27] In our analysis of organizational variation, the parameter of interest was the across-organization standard deviation of the low-value service composite score. To aid in interpretation, we also present a corresponding ratio of the adjusted use of low-value services in an organization at the 90th percentile to that of an organization at the 10th percentile. This measure has been used previously to describe regional variation in health care spending.[12] For the analysis of persistence in organizational behavior, which used a correlated random effects model, the parameter of interest was the correlation coefficient between composite scores in 2010 and 2011. For the analysis of consistency in organizational behavior, which also used a correlated random effects model, the parameters of interest were the pairwise correlations between organizations' different service category composite scores. As a sensitivity analysis, we repeated our analysis of consistency without adjusting for patient HRR. We calculated 95% confidence intervals for estimates via bootstrapping.

Analyses were performed in SAS version 9.3, Stata version 13.1, and R version 3.1.1. Institutional review board approval was obtained through the National Bureau of Economic Research and the Harvard University Faculty of Arts and Sciences.

Table 2.1.1 Measures of Low-Value Services

Clinical Category	Measure	Source	Operational definition	Denominator
Cancer Screening	Cancer screening for patients with CKD receiving dialysis	CW[29]	Screening for cancer of the breast, cervix, colon, or prostate for patients over age 75 with chronic kidney disease receiving dialysis services[a]	Patients with CKD[k] receiving dialysis[l]
	Cervical cancer screening for women age 65 and over	CW USPSTF[30]	Screening Papanicolaou test for women over age 65 with no personal history of cancer or dysplasia noted in claim or in prior claims, and no diagnoses of other female genital cancers, abnormal Papanicolaou findings, or human papillomavirus positivity in prior claims[b]	Women over 65
	Colorectal cancer screening for adults over age 85	USPSTF[31]	Colorectal cancer screening (colonoscopy, sigmoidoscopy, barium enema, or fecal occult blood testing) for patients age 86 or over with no history of colon cancer	Patients over 75
	PSA testing for men age 75 and over	USPSTF[32]	PSA testing for patients age 75 and over with no history of prostate cancer	Men over 75
Diagnostic and Preventive Testing	Bone mineral density testing at frequent intervals	Literature[33,34]	Bone mineral density test within two years of a prior bone mineral density test for patients with an established osteoporosis diagnosis	Patients with osteoporosis[k]
	Homocysteine testing in cardiovascular disease	Literature[35]	Homocysteine testing with no diagnoses of folate or B12 deficiencies in the claim and no folate or B12 testing in prior claims	All patients
	Hypercoagulability testing for patients with deep vein thrombosis	CW[36]	Lab tests for hypercoagulable states within 30 days following diagnosis of lower extremity deep vein thrombosis or pulmonary embolism; no prior evidence of recurrent thrombosis, defined by diagnosis of DVT or pulmonary embolism more than 90 days prior to the testing claim	Patients with deep vein thrombosis[l]
	PTH measurement for patients with stage 1-3 CKD	NICE[37,38]	PTH measurement for patients with chronic kidney disease and no dialysis services before PTH testing or within 30 days following testing, as well as no hypercalcemia diagnosis during the year	Patients with CKD[k] not receiving dialysis[l]

Table 2.1.1 (Continued) Measures of Low-Value Services

Diagnostic and Preventive Testing	Total or free T3 level testing for patients with hypothyroidism	CW[39]	Total or free T3 measurement in a patient with a hypothyroidism diagnosis during the year	Patients with hypothyroidism[l]
	1,25-dihydroxyvitamin D testing in the absence of hypercalcemia or decreased kidney function	CW[40]	Calcitriol testing for patients without hypercalcemia, secondary hyperparathyroidism of renal origin, or conditions related to non-PTH mediated hypercalcemia noted in claim (sarcoidosis, TB, selected neoplasms), and without a history of chronic kidney disease; no diagnosis of hypercalcemia in the past 30 days	All patients
Preoperative Testing	Preoperative chest radiography	CW CADTH[41,42]	Chest x-ray not associated with inpatient or emergency care[c] and occuring within 30 days prior to a low or intermediate risk non-cardiothoracic surgical procedure[d]	Patients undergoing selected surgeries[l]
	Preoperative echocardiography	CW[43]	Echocardiogram not associated with inpatient or emergency care and occuring within 30 days prior to a low or intermediate risk non-cardiothoracic surgical procedure[d]	Patients undergoing selected surgeries[l]
	Preoperative PFT	CW[44]	PFT not associated with inpatient or emergency care and occuring within 30 days prior to a low or intermediate risk surgical procedure[e]	Patients undergoing selected surgeries[l]
	Routine preoperative stress tests	CW[45]	Stress electrocardiogram, echocardiogram, nuclear medicine imaging, cardiac MRI or CT angiography, not associated with inpatient or emergency care and occuring within 30 days prior to a low or intermediate risk surgical procedure[d]	Patients undergoing selected surgeries[l]
Imaging	CT of the sinuses for uncomplicated acute rhinosinusitis	CW[46]	Maxillofacial CT study with a diagnosis of sinusitis and no complications of sinusitis,[f] immune deficiencies, nasal polyps, or head/face trauma noted in claim and no sinusitis diagnosis between 30 and 365 days prior to imaging	All patients

Table 2.1.1 (Continued) Measures of Low-Value Services

Imaging	Head imaging in the evaluation of syncope	CW[47]	CT or MR imaging of the head with a diagnosis of syncope and no diagnoses in claim warranting imaging[g]	Patients with syncope diagnosis[l]
	Head imaging for uncomplicated headache	CW[48]	Brain CT or MR imaging with non-post-traumatic, non-thunderclap headache diagnosis, and no diagnoses in claim warranting imaging[h]	All patients
	EEG for headaches	CW[49]	EEG with headache diagnosis in claim, and no epilepsy or convulsions noted in current or prior claims	All patients
	Back imaging for patients with non-specific low back pain	CW, NICE[50]	Back imaging with a diagnosis of lower back pain occurring within 6 weeks of initial back pain diagnosis and with no indication of radiculopathy or other diagnoses in claim warranting imaging[i]	All patients
	Screening for carotid artery disease in asymptomatic adults	CW, USPSTF[51]	Carotid imaging not associated with inpatient or emergency care for patients without a history of stroke or TIA, and without a diagnosis of stroke, TIA, or focal neurological symptoms in claim	All patients
	Screening for carotid artery disease for syncope	CW[47]	Carotid imaging with syncope diagnosis for patients without a history of stroke or TIA, and without a diagnosis of stroke, TIA, or focal neurological symptoms in claim	Patients with syncope diagnosis[l]
	Imaging for diagnosis of plantar fasciitis	CW[52]	Radiographic or MR imaging with diagnosis of plantar fasciitis occurring within two weeks of initial foot pain diagnosis	Patients with fasciitis diagnosis[l]
Cardiovascular testing and procedures	Stress testing for stable coronary disease	CW[53,54]	Stress testing not associated with inpatient or emergency care[j] for patients with an established diagnosis of acute myocardial infarction (≥ 6 mo before testing)	IHD patients[k]
	Percutaneous coronary intervention with balloon angioplasty or stent placement for stable coronary disease	Literature[54,55]	Coronary stent placement or balloon angioplasty, not associated with an ER visit,[j] for patients with an established diagnosis of acute myocardial infarction (≥ 6 mo before testing)	IHD patients[k]

Table 2.1.1 (Continued) Measures of Low-Value Services

Cardiovascular testing and procedures	Renal artery angioplasty or stenting	Literature[56,57]	Renal/visceral angioplasty or stent placement with a diagnosis of renal atherosclerosis or renovascular hypertension noted in procedure claim	Patients with hypertension[l]
	Carotid endarterectomy for asymptomatic patients	CW[51,58]	Carotid endarterectomy, not associated with an ER visit,j for female patients without a history of stroke or TIA and without stroke, TIA, or focal neurological symptoms noted in claim	All patients
	Inferior vena cava filters for the prevention of pulmonary embolism	Literature[59,60]	Any IVC filter placement	All patients
	Pulmonary Artery Catheterization in the ICU	Literature[61]	Pulmonary artery catheterization for monitoring purposes during an inpatient stay that involved an ICU and a non-surgical DRG; claim contains no diagnoses indicating pulmonary hypertension, cardiac tamponade, or preoperative assessment	Patients who were hospitalized with a non-surgical MS-DRG[l]
Other invasive procedures	Vertebroplasty or kyphoplasty for osteoporotic vertebral fractures	Literature[62–64]	Vertebroplasty/kyphoplasty for vertebral fracture, with no bone cancers, myeloma, or hemangioma noted in procedure claim.	Patients with osteoporosis[k]
	Arthroscopic surgery for knee osteoarthritis	NICE[65]	Arthroscopic debridement/chondroplasty of the knee with diagnosis of osteoarthritis or chondromalacia in the procedure claim and no meniscal tears noted in procedure claim	All patients
	Spinal injection for low-back pain	Literature[66,67]	Outpatient epidural, facet, or trigger point injections for lower back pain, excluding etanercept; no radiculopathy diagnoses in the claim	All patients

Abbreviations: CKD, chronic kidney disease; CT, computed tomography; ED, emergency department; EEG, electroencephalography; ICU, intensive care unit; IVC, inferior vena cava; MR, magnetic resonance; PFT, pulmonary function testing; PSA, prostate-specific antigen; PTH, parathyroid hormone; TIA, transient ischemic attack;

Table 2.1.1 (Continued) Measures of Low-Value Services

[a] The age cutoff is included because transplantation is uncommon in this patient population.

[b] Prior claims refers throughout the table to claims for services before the day of the measured service and during or after the prior calendar year.

[c] Inpatient-associated is defined here as occurring during within 30 days after an inpatient stay; ED-associated, during or 1 day after an ED visit.

[d] Includes breast procedures, colectomy, cholecystectomy, transurethral resection of the prostate, hysterectomy, orthopedic surgical procedures other than hip and knee replacement, corneal transplant, cataract removal, retinal detachment, hernia repair, lithotripsy, arthroscopy, and cholecystectomy.

[e] Includes procedures listed immediately above as well as coronary artery bypass graft, aneurysm repair, thromboendarterectomy, percutaneous transluminal coronary angioplasty, and pacemaker insertion.

[f] Includes inflammation of eyelid or orbit, orbital cellulitis, and visual problems.

[g] Exclusion diagnoses include epilepsy, stroke/TIA, history of stroke, head trauma, convulsions, altered mental status, nervous system symptoms (e.g. hemiplegia), speech problems.

[h] Exclusion diagnoses include those listed immediately above as well as giant cell arteritis, cancer and history of cancer.

[i] Exclusion diagnoses include cancer, trauma, intravenous drug abuse, neurological impairment, endocarditis, septicemia, tuberculosis, osteomyelitis, fever, weight loss, loss of appetite, night sweats, and anemia.

[j] Inpatient-associated is defined here as occurring during an inpatient stay; ED-associated, during or within 14 d after an ED visit.

[k] Defined by the presence of CCW first indication date prior to December 31st of the year.

[l] Defined by presence of relevant diagnosis or service codes during the year.

2.1.3 RESULTS

The general sample consisted of 4,039,733 beneficiaries attributed to 3,137 organizations. The ACO sample consisted of 1,432,644 beneficiaries attributed to 250 ACOs. Beneficiary characteristics were largely similar between the two samples (Table 2.1.2). Organizations in the ACO sample were considerably larger, with an average of 5,731attributed beneficiaries compared to 1,288 in the general sample. On average, organizations in the general sample delivered an unadjusted rate of 45.6 low-value services per 100 beneficiaries and those in the ACO sample delivered 47.7 services per 100 beneficiaries. Standard deviations of 13.8 and 11.4 services per 100 beneficiaries, respectively, suggest substantial variation in unadjusted low-value service delivery. Low-value imaging and low-value cancer screening were more frequent than the other service categories.

Table 2.1.3 presents adjusted estimates of variation in organizations' delivery of low-value services. The across-organization standard deviation in the use of low-value services was 9.3 services per 100 beneficiaries in the general sample (95% CI 8.8–9.9) and 7.6 services per 100 beneficiaries in the ACO sample (95% CI 6.8–8.3), without adjustment for geographic region. This corresponds to 90th/10th percentile ratios of 1.71 (95% CI 1.65–1.77) and 1.51 (95% CI 1.45–1.58), respectively. Models that adjusted for geographic region produced smaller estimates of variation, with 90th/10th percentile ratios of 1.51 (95% CI 1.46–1.54) in the general sample and 1.27 (95% CI 1.23–1.32) in the ACO sample. Organizations' adjusted low-value service use was highly persistent, with correlation coefficients between 2010 and 2011 service use of 0.95 (95% CI 0.92–0.98) in the general sample and 0.87 (95% CI 0.80–0.95) in the ACO sample.

Within organizations, alternate categories of low-value services were positively correlated with one another in (Table 2.1.4). Adjusted correlations between categories were positive and statistically significant for 13/15 pairs of services in the general sample and 10/15 pairs of services in the ACO sample. All non-significant correlations involved a single category of low-value services, other invasive procedures. The average correlation coefficient across all pairs was 0.19 in the general sample (95% CI 0.17–0.21) and 0.24 in the ACO sample (95% CI 0.16–0.32). The corresponding averages were 0.23 and 0.34 among the pairs that did not include other invasive procedures. In both samples, the greatest correlation was between low-value cardiovascular testing and procedures and low-value imaging. Low-value imaging had the highest correlation with other categories, with an average of 0.29 in the general sample and 0.39 in the ACO sample. In a sensitivity analysis without adjustment for patient region (Appendix 2), this pattern of correlations was broadly similar. However, the average correlation between service categories increased to 0.23 in the general sample (95% CI 0.21–0.25) and 0.35 in the ACO sample (95% CI 0.28–0.41).

Table 2.1.2. Unadjusted Beneficiary and Provider Organization Characteristics

	General Sample	ACO Sample
Beneficiaries, no.	4,039,733	1,432,644
Observations, no. of beneficiary-years	10,149,111	3,580,702
Mean age, y	72.6 ± 11.7	72.4 ± 11.9
Female sex, %	58.4	58.2
Race/ethnicity, %		
White	89.4	87.3
Black	7.4	7.2
Hispanic	0.9	1.7
Other	2.2	3.8
Medicaid recipient, %	16.7	20.8
Disabled,[a] %	20.3	21.1
End-stage renal disease, %	1.0	1.1
CCW conditions[b]		
Total no., mean	5.1 ± 2.6	5.2 ± 2.7
≥6 conditions	43.8	46.1
≥9 conditions	16.0	17.8
Low-value service measure denominators qualified for, mean	14.9 ± 2.5	14.9 ± 2.5
ZCTA characteristics for age 65+, mean		
median income	38,801 ± 13,533	39,706 ± 14,161
% below FPL	8.5	8.4
% with high school degree	77.2	77.3
% with college degree	20.3	21.1
	3,137	250
Beneficiaries per organization, mean no.	1,288 ± 1,447	5,731 ± 5,135
Low-value services per 100 beneficiaries, mean		
Cancer screening	14.0 ± 5.6	14.0 ± 3.8
Diagnostic and Preventive Testing	7.8 ± 0.6	9.2 ± 4.2
Preoperative Testing	2.5 ± 1.2	2.6 ± 1.0
Imaging	15.7 ± 5.2	16.3 ± 4.0
Cardiovascular testing and procedures	1.1 ± 0.6	1.2 ± 0.4
Other invasive procedures	4.4 ± 2.2	4.4 ± 1.6
Total	45.6 ± 13.8	47.7 ± 11.4

ACO = Accountable Care Organization, CCW = Chronic Conditions Warehouse, HCC = Hierarchical Condition Categories, ZCTA = ZIP Code Tabulation Area. Estimates are derived from 2007-2011 data. All means and percentages are unadjusted. Means are presented ± standard deviations.

[a] Refers to beneficiaries for whom disability was the original reason for Medicare eligibility.

[b] Chronic conditions include 25 conditions from the CCW: acute myocardial infarction, Alzheimer's disease, Alzheimer's disease and related disorders or senile dementia, anemia, asthma, atrial fibrillation, benign prostatic hyperplasia, breast cancer, chronic kidney disease, chronic obstructive pulmonary disease, colorectal cancer, depression, diabetes, endometrial cancer, heart failure, hip/pelvic fracture, hyperlipidemia, hypertension, hypothyroidism, ischemic heart disease, lung cancer, osteoporosis, prostate cancer, rheumatoid arthritis/osteoarthritis, stroke/transient ischemic attack.

Table 2.1.3. Variation and Persistence of Low-Value Service Delivery

General Sample

Variation in Low-Value Services Per 100 Beneficiaries

	Standard Deviation	95% CI	90th/10th Percentile Ratio	95% CI
Without Adjustment for Region	9.3	8.8—9.9	1.71	1.65—1.77
With Adjustment for Region	7.1	6.6—7.6	1.50	1.46—1.54

Persistence of Low-Value Services Per 100 Beneficiaries

Correlation Between Years	95% CI
0.95	0.92—0.98

ACO Sample

Variation in Low-Value Services Per 100 Beneficiaries

	Standard Deviation	95% CI	90th/10th Percentile Ratio	95% CI
Without Adjustment for Region	7.6	6.8—8.3	1.51	1.45—1.58
With Adjustment for Region	4.4	3.8—5.1	1.27	1.23—1.32

Persistence of Low-Value Services Per 100 Beneficiaries

Correlation Between Years	95% CI
0.87	0.80—0.95

TIN = Taxpayer Identification Number, ACO = Accountable Care Organization

Estimates are derived from models of organizations' total composite low-value service use. Standard deviation estimates refer to across-organization variation, estimated via a random intercept model that includes 2007-2011 data. Correlation between years refers to the correlation coefficient for organizational performance between organizations' 2010 and 2011 performance. Models adjust for overdispersion of observed performance, beneficiary sociodemographic and clinical characteristics, local area economic characteristics, patient region, year, and the number of patients qualifying for component measure denominators.

Table 2.1.4 Consistency Across Low-Value Service Domains Within Provider Organizations

General Sample

Measure composite	Cancer Screening	Diag.	Preop.	Imaging	Cardio.	Other.
Cancer screening	-	-	-	-	-	-
Diagnostic and Preventive Testing	0.19 (0.13—0.25)	-	-	-	-	-
Preoperative Testing	0.16 (0.10—0.21)	0.17 (0.10—0.24)	-	-	-	-
Imaging	0.33 (0.27—0.38)	0.20 (0.15—0.26)	0.29 (0.24—0.34)	-	-	-
Cardiovascular testing and procedures	0.12 (0.05—0.18)	0.18 (0.11—0.24)	0.26 (0.19—0.32)	0.44 (0.37—0.51)	-	-
Other invasive procedures	0.16 (0.11—0.22)	0.04 (-0.02—0.10)	0.01 (-0.06—0.08)	0.18 (0.12—0.24)	0.12 (0.04—0.19)	-

ACO Sample

Measure composite	Cancer Screening	Diag.	Preop.	Imaging	Cardio.	Other.
Cancer screening	-	-	-	-	-	-
Diagnostic and Preventive Testing	0.16 (-0.01—0.33)	-	-	-	-	-
Preoperative Testing	0.30 (0.13—0.47)	0.17 (0.04—0.30)	-	-	-	-
Imaging	0.42 (0.28—0.56)	0.40 (0.23—0.57)	0.42 (0.26—0.57)	-	-	-
Cardiovascular testing and procedures	0.24 (0.06—0.42)	0.28 (0.08—0.48)	0.51 (0.36—0.67)	0.53 (0.35—0.71)	-	-
Other invasive procedures	0.08 (-0.07—0.24)	-0.03 (-0.23—0.17)	-0.08 (-0.27—0.11)	0.16 (-0.03—0.36)	0.03 (-0.22—0.28)	-

ACO = Accountable Care Organization

Estimates are correlation coefficients and their 95% confidence intervals. For each sample of organizations, all correlations are derived from a single model of organizations' low-value service composite measures. The models adjust for overdispersion of observed performance, beneficiary sociodemographic and clinical characteristics, local area economic characteristics, patient region, year, and the number of patients qualifying for component measure denominators.

2.1.4 DISCUSSION

In this national study of low-value service use among provider organizations, we observed substantial variation in service use across organizations, highly persistent service use over time, and positive associations between various components of low-value care. Variation in organizations' use of low-value services was large compared to previous estimates of practice pattern variation. For example, the 90th/10th percentile ratio of regions' Medicare adjusted spending has been estimated at 1.25,[12] smaller than our estimates of variation among organizations. We also observed that organizations varied substantially within the same region, which is consistent with prior studies of within-region variation in overall spending.[28] We observed less variation in the ACO sample than the general sample, which may reflect greater uniformity among larger provider organizations in general or among organizations opting to participate in the ACO program specifically. Despite the considerable variation across-organizations, low-value services were used frequently even among the best performing organizations.

Positive correlations between different clinical categories of low-value services suggest some consistency of low-value service delivery within an organization. However, because these associations were generally modest, a single category of low-value service measures cannot provide a precise prediction of low-value service use in other clinical domains. Profiling organizations on the basis of many low-value practices may produce more reliable estimates, as evidenced by high correlations between organizations' composite scores over time. Somewhat greater correlations for ACOs than for organizations in the general sample suggests that ACOs may have more effective mechanisms in place for influencing providers than other organizations. However, even for ACOs, we observed only weak associations between invasive

procedures, which tend to be provided by surgeons, and other services. This finding suggests that different physician specialties within an organization may have weakly associated or independent practice styles with respect to low-value services.

There are several limitations to this study. First, it is a descriptive analysis, and the results may diverge from the outcomes that would be observed if patients and physicians were randomized to different provider organizations. Second, although we observed variation in organization's use of low-value services, we could not isolate which characteristics of organizations drove this variation. Third, although our analyses adjust for many patient sociodemographic and clinical characteristics, it is possible that unobserved differences in patient characteristics may have contributed to organizations' measured low-value service use. Fourth, because some measures of low-value services depend on diagnostic codes in claims, variation in measured service use might reflect differences in the completeness of organizations' diagnostic coding. However, many services we examined were not detected on the basis of diagnoses in claims. Fifth, some of the organizations in our sample, especially smaller organizations, may not have delivered all of the types of services that we measured. This was one motivation for examining ACOs, which are more likely to be a part of integrated delivery systems that provide a comprehensive range of services.

Our study indicates that organizations have exhibited distinct profiles in their use of low-value services, with substantial performance variations that are persistent over time. Our methods of characterizing organizations' use of low-value service may have applications for public disclosure (e.g. organization report cards) or benefit design (e.g. tiering of organizations), or for future efforts to study the drivers of organizational variation. Our findings are consistent with, though not definitive evidence of,

organizations influencing the amounts of low-value services that their patients receive. If organizations do shape the practice patterns of their affiliated physicians, then policies like global payment contracts, which modify organization-level incentives, are more likely to affect patient care.

2.1.5 REFERENCES

1. Song Z, Rose S, Safran DG, Landon BE, Day MP, Chernew ME. Changes in health care spending and quality 4 years into global payment. *New England Journal of Medicine*. 2014;371(18):1704–14.
2. McWilliams JM, Landon BE, Chernew ME, Zaslavsky AM. Changes in patients' experiences in Medicare Accountable Care Organizations. *New England Journal of Medicine*. 2014;371(18):1715–24.
3. Rebitzer JB, Votruba M. Organizational Economics and Physician Practices. In: Culyer AJ, ed. *Encyclopedia of Health Economics*. Volume 2. Newnes; 2014:414–424.
4. Robinson JC. Theory and Practice in the Design of Physician Payment Incentives. *The Milbank Quarterly*. 2001;79(2):149–177.
5. Bright TJ, Wong A, Dhurjati R, et al. Effect of clinical decision-support systems: a systematic review. *Annals of Internal Medicine*. 2012;157(1):29–43.
6. Javitt JC, Rebitzer JB, Reisman L. Information technology and medical missteps: evidence from a randomized trial. *Journal of Health Economics*. 2008;27(3):585–602.
7. Jackson GL, Powers BJ, Chatterjee R, et al. The Patient-Centered Medical Home: A Systematic Review. *Annals of Internal Medicine*. 2013;158(3):169.
8. Landon BE, Keating NL, Barnett ML, et al. Variation in Patient-Sharing Networks of Physicians Across the United States. *JAMA*. 2012;308(3).
9. Asch DA, Nicholson S, Srinivas S. Evaluating Obstetrical Residency Programs Using Patient Outcomes. *JAMA*. 2010.
10. Asch D a, Epstein A, Nicholson S. Evaluating medical training programs by the quality of care delivered by their alumni. *JAMA*. 2007;298(9):1049–51.

11. Doyle JJ, Ewer SM, Wagner TH. Returns to Physician Human Capital: Evidence from Patients Randomized to Physician Teams. *Journal of Health Economics*. 2010;29(6):866–82.
12. Institute of Medicine. *Variation in Health Care Spending: Target Decision Making, Not Geography*. (Newhouse JP, Garber AM, Graham RP, McCoy MA, Mancher M, Kibria A, eds.).; 2013.
13. Jha AK, Orav EJ, Zheng J, Epstein AM. Patients' perception of hospital care in the United States. *The New England Journal of Medicine*. 2008;359(18):1921–31.
14. Isaac T, Zaslavsky AM, Cleary PD, Landon BE. The relationship between patients' perception of care and measures of hospital quality and safety. *Health Services Research*. 2010;45(4):1024–40.
15. Cassel CK, Guest JA. Choosing wisely: helping physicians and patients make smart decisions about their care. *JAMA*. 2012;307(17):1801–2.
16. Welch WP, Cuellar AE, Stearns SC, Bindman AB. Proportion of physicians in large group practices continued to grow in 2009-11. *Health Affairs*. 2013;32(9):1659–66.
17. Guidestar. Available at: http://www.guidestar.org/. Accessed November 20, 2014.
18. EINFinder. Employer Identification Number Database. Available at: http://www.einfinder.com/. Accessed November 20, 2014.
19. McWilliams JM, Chernew ME, Landon BE, Schwartz AL. Performance differences in year 1 of Pioneer accountable care organizations. 2015;Submitted.
20. Schwartz AL, Landon BE, Elshaug AG, Chernew ME, McWilliams JM. Measuring low-value care in Medicare. *JAMA Internal Medicine*. 2014;174(7):1067–76.

21. Choosing Wisely. Lists of five things physicians and patients should question. 2014. Available at: http://www.choosingwisely.org/doctor-patient-lists/. Accessed February 16, 2015.
22. U.S. Preventive Services Task Force. Published Recommendations. 2014. Available at: www.uspreventiveservicestaskforce.org/BrowseRec/Index/brouse-recommendations. Accessed February 16, 2015.
23. Canadian Agency for Drugs and Technologies in Health. Health technology assessments. Available at: http://cadth.ca/en/products/health-technology-assessment. Accessed February 16, 2015.
24. Elshaug AG, Moss JR, Littlejohns P, Karnon J, Merlin TL, Hiller JE. Identifying existing health care services that do not provide value for money. *Medical Journal of Australia*. 2012;190(5):269–73.
25. The Dartmouth Atlas of Health Care. Available at: http://www.dartmouthatlas.org/. Accessed May 20, 2014.
26. Agency for Healthcare Research and Quality. Instructions for Analyzing Data from CAHPS Surveys. 2015. Available at: https://cahps.ahrq.gov/surveys-guidance/survey4.0-docs/2015-Instructions-for-Analyzing-Data-from-CAHPS-Surveys.pdf.
27. Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. *International Journal for Quality in Health Care*. 2001;13(6):481–8.
28. Zhang Y, Baik SH, Fendrick AM, Baicker K. Comparing local and regional variation in health care spending. *New England journal of medicine*. 2012;367(18):1724–31.
29. Holley JL. Screening, diagnosis, and treatment of cancer in long-term dialysis patients. *Clinical Journal of the American Society of Nephrology*. 2007;2(3):604–10.

30. Vesco K, Whitlock E, Eder M, et al. *Screening for cervical cancer: a systematic evidence review for the U.S. Preventive Services Task Force*. (Evidence . Rockville, MD: Agency for Healthcare Research and Quality; 2011.
31. Whitlock EP, Lin J, Liles E, Beil T, Fu R. Screening for colorectal cancer: A targeted, updated systematic review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*. 2008;149(9):638.
32. Lin K, Lipsitz R, Miller T, Janakiraman S. Benefits and harms of prostate-specific antigen screening for prostate cancer: an evidence update for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*. 2008;149(3):192.
33. Bell KJL, Hayen A, Macaskill P, et al. Value of routine monitoring of bone mineral density after starting bisphosphonate treatment: secondary analysis of trial data. *BMJ*. 2009;338.
34. Hillier TA, Stone KL, Bauer DC, et al. Evaluating the value of repeat bone mineral density measurement and prediction of fractures in older women: the study of osteoporotic fractures. *Archives of Internal Medicine*. 2007;167(2):155–60.
35. Martí-Carvajal AJ, Solà I, Lathyris D, Karakitsiou D-E, Simancas-Racines D. Homocysteine-lowering interventions for preventing cardiovascular events. *The Cochrane database of systematic reviews*. 2013;1:CD006612.
36. Baglin T, Gray E, Greaves M, et al. Clinical guidelines for testing for heritable thrombophilia. *British Journal of Haematology*. 2010;149(2):209–20.
37. Levin A, Bakris GL, Molitch M, et al. Prevalence of abnormal serum vitamin D, PTH, calcium, and phosphorus in patients with chronic kidney disease: results of the study to evaluate early kidney disease. *Kidney International*. 2007;71(1):31–8.

38. Palmer SC, McGregor DO, Craig JC, Elder G, Macaskill P, Strippoli GF. Vitamin D compounds for people with chronic kidney disease not requiring dialysis. *The Cochrane Database of Systematic Reviews*. 2009;(4):CD008175.
39. Garber JR, Cobin RH, Gharib H, et al. Clinical practice guidelines for hypothyroidism in adults: cosponsored by the American Association of Clinical Endocrinologists and the American Thyroid Association. *Endocrine Practice*. 2012;18(6):988–1028.
40. Holick MF. Vitamin D deficiency. *New England Journal of Medicine*. 2007;357(3):266–81.
41. Mohammed T, Kirsch J, Amorosa J, et al. *ACR Appropriateness Criteria routine admission and preoperative chest radiography*. Reston, VA; 2011.
42. Joo HS, Wong J, Naik VN, Savoldelli GL. The value of screening preoperative chest x-rays: a systematic review. *Canadian Journal of Anaesthesia*. 52(6):568–74.
43. Douglas PS, Garcia MJ, Haines DE, et al. CF/ASE/AHA/ASNC/HFSA/HRS/SCAI/ SCCM/SCCT/SCMR 2011 appropriate use criteria for echocardiography. *Journal of the American College of Cardiology*. 2011;57(9):1126–66.
44. Qaseem A, Snow V, Fitterman N, et al. Risk assessment for and strategies to reduce perioperative pulmonary complications for patients undergoing noncardiothoracic surgery: a guideline from the American College of Physicians. *Annals of Internal Medicine*. 2006;144(8):575–80.
45. Fleisher LA, Beckman JA, Brown KA, et al. ACC/AHA 2007 guidelines on perioperative cardiovascular evaluation and care for noncardiac surgery. *Circulation*. 2007;116(17):e418–99.
46. Cornelius RS, Martin J, Wippold FJ, et al. ACR appropriateness criteria sinonasal disease. *Journal of the American College of Radiology*. 2013;10(4):241–6.

47. Moya A, Sutton R, Ammirati F, et al. Guidelines for the diagnosis and management of syncope (version 2009). *European Heart Journal*. 2009;30(21):2631–71.
48. Jordan J, Wippold FI, Cornelius R, et al. *ACR appropriateness criteria headache*. Reston, VA; 2009.
49. Gronseth GS, Greenberg MK. The utility of the electroencephalogram in the evaluation of patients presenting with headache: a review of the literature. *Neurology*. 1995;45(7):1263–7.
50. Chou R, Fu R, Carrino JA, Deyo RA. Imaging strategies for low-back pain: systematic review and meta-analysis. *Lancet*. 2009;373(9662):463–72.
51. Wolff T, Guirguis-Blake J, Miller T, Gillespie M, Harris R. Screening for carotid artery stenosis: an update of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*. 2007;147(12):860–70.
52. Buchbinder R. Plantar fasciitis. *New England Journal of Medicine*. 2004;350(21):2159–66.
53. Hendel RC, Berman DS, Di Carli MF, et al. ACCF/ASNC/ACR/AHA/ASE/SCCT/ SCMR/SNM 2009 appropriate use criteria for cardiac radionuclide imaging. *Circulation*. 2009;119(22):e561–87.
54. Boden WE, O'Rourke RA, Teo KK, et al. Optimal medical therapy with or without PCI for stable coronary disease. *New England Journal of Medicine*. 2007;356(15):1503–16.
55. Lin GA, Dudley RA, Redberg RF. Cardiologists' use of percutaneous coronary interventions for stable coronary artery disease. *Archives of Internal Medicine*. 2007;167(15):1604–9.

56. Wheatley K, Ives N, Gray R, et al. Revascularization versus medical therapy for renal-artery stenosis. *New England Journal of Medicine*. 2009;361(20):1953–62.
57. Cooper CJ, Murphy TP, Cutlip DE, et al. Stenting and medical therapy for atherosclerotic renal-artery stenosis. *New England Journal of Medicine*. 2014;370(1):13–22.
58. Goldstein LB, Bushnell CD, Adams RJ, et al. Guidelines for the primary prevention of stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2011;42(2):517–84.
59. PREPIC Study Group. Eight-year follow-up of patients with permanent vena cava filters in the prevention of pulmonary embolism: the PREPIC (Prevention du Risque d'Embolie Pulmonaire par Interruption Cave) randomized study. *Circulation*. 2005;112(3):416–22.
60. Sarosiek S, Crowther M, Sloan JM. Indications, complications, and management of inferior vena cava filters: the experience in 952 patients at an academic hospital with a level I trauma center. *JAMA Internal Medicine*. 2013;173(7):513–7.
61. Rajaram SS, Desai NK, Kalra A, et al. Pulmonary artery catheters for adult patients in intensive care. *The Cochrane Database of Systematic Reviews*. 2013;2.
62. Kallmes DF, Comstock BA, Heagerty PJ, et al. A randomized trial of vertebroplasty for osteoporotic spinal fractures. *The New England Journal of Medicine*. 2009;361(6):569–79.
63. Buchbinder R, Osborne RH, Ebeling PR, et al. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. *New England Journal of Medicine*. 2009;361(6):557–68.

64. Boonen S, Van Meirhaeghe J, Bastian L, et al. Balloon kyphoplasty for the treatment of acute vertebral compression fractures: 2-year results from a randomized trial. *Journal of Bone and Mineral Research*. 2011;26(7):1627–37.

65. Laupattarakasem W, Laopaiboon M, Laupattarakasem P, Sumananont C. Arthroscopic Debridement for Knee Osteoarthritis. *Cochrane Database of Systematic Reviews*. 2008;(1):CD005118.

66. Pinto RZ, Maher CG, Ferreira ML, et al. Epidural corticosteroid injections in the management of sciatica: a systematic review and meta-analysis. *Annals of Internal Medicine*. 2012;157(12):865–77.

67. Staal JB, de Bie R, de Vet HC, Hildebrandt J, Nelemans P. Injection therapy for subacute and chronic low-back pain. *The Cochrane Database of Systematic Reviews*. 2008;(3):CD001824.

2.2 Changes in Low-Value Services in Year 1 of the Medicare Pioneer ACO Program

2.2.1 Introduction

Reducing unnecessary health care utilization, a source of substantial spending,[1] has been a central goal of many government[2–4] and private initiatives.[5,6] However, distinguishing high-value from low-value use of the same service is often challenging. As a result, efforts to directly limit overuse of specific services through coverage restrictions or other payment incentives may produce unintended consequences or achieve minimal gains.[7–9] Another approach to enhancing value is to place spending for all services under a global budget, with incentives to stay under the budget and to improve performance on quality measures, as in the Medicare Pioneer accountable care organization (ACO) program. This approach has been associated with lower overall spending and improved or stable performance on standard quality measures.[10–14]

However, it is unknown whether payment reforms like the Pioneer ACO program are associated with reductions in overuse. A combination of lower overall spending and improved performance on quality measures can result from reductions in high-value services affecting unmeasured dimensions of quality rather than from reductions in low-value services. Also, because risk-based contracts do not incentivize reductions in overuse directly, it is unclear whether providers under these contracts are targeting low-value services in their broader efforts to control overall spending. If ACO-like payment models succeed in reducing wasteful utilization, there should be observable reductions in the delivery of low-value services that can be measured directly. Moreover, if providers are targeting low-value services specifically in response

to ACO contracts, their efforts should result in greater reductions in spending on low-value services than in overall spending.

We constructed 31 claims-based measures of low-value services—services that provide minimal clinical benefit on average. Using these measures and 2009-2012 Medicare fee-for-service (FFS) claims, we conducted a difference-in-differences analysis comparing use of low-value services between beneficiaries served by Pioneer ACOs and beneficiaries served by non-ACO providers before vs. after the start of Pioneer contracts in 2012.

2.2.2 METHODS

Background on the Pioneer ACO Program

In 2012, 32 provider organizations volunteered to participate in the Medicare Pioneer ACO program, in which each ACO receives a bonus payment, or is penalized, if overall spending for an attributed patient population falls sufficiently below or above a financial benchmark, respectively. The financial benchmark is based on baseline spending for each ACO's attributed population, inflated each year according to national spending growth. Performance on 33 quality measures determines the proportion of savings or losses shared by the ACO, although ACOs were only required to report on these measures to be eligible for maximum savings in the first year of the program. None of the quality measures in Medicare ACO contracts assesses overuse of medical services.

Study Population

We analyzed 2009-2012 Medicare claims for a random 20% sample of beneficiaries; in a given year, this sample includes sample members from the prior year plus a 20% sample of new beneficiaries. For each year of the study period, we included beneficiaries in the study sample if they were continuously enrolled in Parts A and B of traditional Medicare while alive during that year and the entire prior year. We used the prior year of claims to collect diagnoses and procedures used for case-mix adjustment or for assessing the appropriateness of service use. In each study year, beneficiaries were excluded if they did not receive primary care services necessary for attribution to provider organizations, or if they were attributed to any of the 114 organizations that entered the Medicare Shared Savings Program (MSSP) later in 2012. MSSP ACOs faced weaker incentives than Pioneer ACOs to reduce spending, and for only part of 2012. Thus, if MSSP ACOs took early steps to limit low-value services, inclusion of their beneficiaries in the control group could have biased our estimates.

Each of the 32 organizations that entered the Pioneer ACO program was defined as the collection of National Provider Identifiers (NPIs) for physicians listed by the ACO as participating in the ACO contract. Our definition of ACOs as sets of NPIs reflects the organizations' ability to include only a subset of their affiliated physicians in their ACO contracts. Following the MSSP attribution rules and previously described methods, [14] for each year in the study period, each beneficiary was assigned to the ACO (ACO group) or non-ACO practice (control group) that accounted for the greatest fraction of that beneficiary's annual allowed charges for primary care services (Appendix 2). Non-ACO practices were defined by taxpayer identification numbers (TINs), which identify the billing practice, provider organization, or individual physician.

Measures of Low-Value Services

We constructed 31 claims-based measures of low-value care, including 26 analyzed in a prior study [8] The selection and construction of these measures is described in Section 2.1.

The primary outcome of this study was use of low-value services, defined as the annual count of all measured services. We chose this primary outcome because measures of overall use provide equal weight to each clinical decision, while measures of spending will tend to be driven by more expensive services. To compare changes in low-value services to previously published estimates of overall spending changes associated with the Pioneer program, we examined price-standardized spending on measured services as a secondary outcome. Methods of standardizing service prices for spending calculations are presented in the Appendix 2.

To assess whether any changes in low-value service use associated with Pioneer ACO contracts were concentrated in a specific clinical area or evident in multiple areas, we categorized the 31 low-value services into the following clinical categories: cancer screening, diagnostic and preventive testing, preoperative testing, imaging, cardiovascular testing and procedures, and other invasive procedures. We also categorized services as higher-priced (standardized price $180-$13,331) or lower-priced ($5-$117) than to the median service price, because ACOs would be unlikely to reduce higher-priced services in the absence of new payment incentives, whereas ACOs might restrict provision of lower-priced wasteful services even under FFS incentives to improve quality without major reductions in revenue. Thus, reductions in use of higher-priced low-value services would provide stronger evidence of changes related specifically to ACO contract incentives.

Finally, to explore the possibility that patient preferences moderated providers' responses to ACO contracts, we categorized services as less vs. more sensitive to patient preferences (Table 2.2.1). For example, we considered testing for hypercoagulability for patients with deep venous thrombosis as less sensitive to patient preferences because most patients would be unaware that such testing could be done. Table 2.1.1 presents each measure's source and supporting literature, operational definition, and assigned categories of price and preference sensitivity, as well as the mean annual count of each service per beneficiary in the pre-contract period.

Table 2.2.1 Summary of Low-Value Care Measures

Clinical Category	Measure	Price category	Preference sensitivity category	Mean annual count per 100 benes (2009-2011)
Cancer Screening	Cancer screening for patients with CKD receiving dialysis	Lower priced	More sensitive	0.1
	Cervical cancer screening for women age 65 and over	Lower priced	More sensitive	4.3
	Colorectal cancer screening for adults over age 85	Lower priced	More sensitive	0.6
	PSA testing for men age 75 and over	Lower priced	More sensitive	7.7
Diagnostic and Preventive Testing	Bone mineral density testing at frequent intervals	Lower priced	Less sensitive	0.6
	Homocysteine testing in cardiovascular disease	Lower priced	Less sensitive	0.8
	Hypercoagulability testing for patients with DVT	Lower priced	Less sensitive	0.04
	PTH measurement for patients with stage 1-3 CKD	Lower priced	Less sensitive	3.8
	Total or free T3 level testing for patients with hypothyroidism	Lower priced	Less sensitive	2.6
	1,25-dihydroxyvitamin D testing in the absence of hypercalcemia or decreased kidney function	Lower priced	Less sensitive	1.0
Pre-operative Testing	Preoperative chest radiography	Lower priced	Less sensitive	1.7
	Preoperative echocardiography	Higher priced	Less sensitive	0.3
	Preoperative PFT	Lower priced	Less sensitive	0.1
	Routine preoperative stress tests	Higher priced	Less sensitive	0.3
Imaging	CT of the sinuses for uncomplicated acute rhinosinusitis	Higher priced	More sensitive	0.3
	Head imaging in the evaluation of syncope	Higher priced	More sensitive	1.0
	Head imaging for uncomplicated headache	Higher priced	More sensitive	2.9
	EEG for headaches	Higher priced	Less sensitive	0.05
	Back imaging for patients with non-specific low back pain	Lower priced	More sensitive	4.4
	Screening for carotid artery disease in asymptomatic adults	Higher priced	Less sensitive	5.8
	Screening for carotid artery disease for syncope	Higher priced	Less sensitive	0.6
	Imaging for diagnosis of plantar fasciitis	Lower priced	More sensitive	0.4
Cardiovascular testing and procedures	Stress testing for stable coronary disease	Higher priced	More sensitive	0.7
	Percutaneous coronary intervention with balloon angioplasty or stent placement for stable coronary disease	Higher priced	More sensitive	0.1
	Renal artery angioplasty or stenting	Higher priced	Less sensitive	0.1

Table 2.2.1 (Continued) Summary of Low-Value Care Measures

Cardiovascular testing and procedures	Carotid endarterectomy for asymptomatic patients	Higher priced	Less sensitive	0.05
	Inferior vena cava filters for the prevention of pulmonary embolism	Higher priced	Less sensitive	0.2
	Pulmonary Artery Catheterization in the ICU	Lower priced	Less sensitive	0.01
Other invasive procedures	Vertebroplasty or kyphoplasty for osteoporotic vertebral fractures	Higher priced	Less sensitive	0.3
	Arthroscopic surgery for knee osteoarthritis	Higher priced	More sensitive	0.2
	Spinal injection for low-back pain	Higher priced	More sensitive	4.0

Abbreviations: CKD, chronic kidney disease; CT, computed tomography; DVT, deep vein thrombosis; ED, emergency department; EEG, electroencephalography; ICU, intensive care unit; IVC, inferior vena cava; MR, magnetic resonance; PFT, pulmonary function testing; PSA, prostate-specific antigen; PTH, parathyroid hormone; TB, tuberculosis; TIA, transient ischemic attack; USPSTF, US Preventive Services Task Force D recommendations.

[a] The age cutoff is included because transplantation is uncommon in this patient population.

[b] Prior claims refers throughout the table to claims for services before the day of the measured service and during or after the prior calendar year.

[c] Inpatient-associated is defined here as occurring during within 30 d after an inpatient stay; ED-associated, during or 1 d after an ED visit.

[d] Includes breast procedures, colectomy, cholecystectomy, transurethral resection of the prostate, hysterectomy, orthopedic surgical procedures other than hip and knee replacement, corneal transplant, cataract removal, retinal detachment, hernia repair, lithotripsy, arthroscopy, and cholecystectomy.

[e] Includes procedures listed immediately above as well as coronary artery bypass graft, aneurysm repair, thromboendarterectomy, percutaneous transluminal coronary angioplasty, and pacemaker insertion.

[f] Includes inflammation of eyelid or orbit, orbital cellulitis, and visual problems.

[g] Exclusion diagnoses include epilepsy, stroke/TIA, history of stroke, head trauma, convulsions, altered mental status, nervous system symptoms (e.g. hemiplegia), speech problems.

[h] Exclusion diagnoses include those listed immediately above as well as giant cell arteritis, cancer and history of cancer.

[i] Exclusion diagnoses include cancer, trauma, intravenous drug abuse, neurological impairment, endocarditis, septicemia, tuberculosis, osteomyelitis, fever, weight loss, loss of appetite, night sweats, and anemia.

[j] Inpatient-associated is defined here as occurring during an inpatient stay; ED-associated, during or within 14 d after an ED visit.

Covariates

For each beneficiary, the following demographic and clinical covariates were assessed from Medicare claims and enrollment files: age (<65, 65-69, 70-74…, >84), sex, race/ethnicity, disability as the original reason for Medicare entitlement, presence of end-stage renal disease, presence of 27 chronic conditions in the Chronic Condition Warehouse (CCW) by the start of each study year (including indicators for each condition and indicators for having ≥2, 3, 4, etc. conditions up to ≥9), and the patient's hierarchical condition category (HCC) risk score. Because most low-value service measures do not apply to all beneficiaries (e.g. low-value PSA tests were defined as PSA tests for men age 75 and over), we also created indicators for whether beneficiaries qualified for potential receipt of each low-value service (see Appendix 2 for definitions of these qualifying indicators).

ACO Baseline Levels of Low-Value Services

Because organizations with more wasteful practices may have a greater opportunity to limit wasteful care, we measured ACO baseline levels of low-value service use and tested whether those levels were associated with changes in the use of low-value services after ACO contacts began. We decomposed an ACO's baseline levels of low-value care into two components. First, we assessed whether the ACO had a greater or lesser risk-adjusted count of low-value services per beneficiary than the control group within an ACO's service area (Appendix 2). Second, we assessed whether the risk-adjusted count of low-value services among the control group in each ACO's service area was greater or less than that of the median ACO's service area. This decomposition allowed us to examine whether an organization's prior performance

relative to its service area or service area performance relative to the national average predicted changes under ACO contracts. This distinction bears on whether ACO contracts might be associated with convergence in provider practices within regions or across regions. Baseline levels of low-value care were assessed in 2008 in order to avoid results driven by regression to the mean between the pre-contract period (2009-2011) and 2012; we found no evidence of regression to the mean over the 2009-2011 pre-contract period (Appendix 2).

Statistical Analysis

We conducted a difference-in-differences analysis to quantify changes in the annual per-beneficiary count of low-value services in the ACO group that differed from concurrent changes in the control group from the pre-contract period (2009-2011) to the post-contract period (2012), while adjusting for geography and any coincident changes in the groups' measured patient characteristics. Specifically, we fit the following linear regression model:

$$E(Y_{itkh}) = \beta_0 + \beta_1 ACO\_indicators_k + \beta_2 HRR\_indicators_h \times Year_t + \beta_3 ACO\_contract_{kt} + \beta_4 Covariates_{it}$$

with $E(Y_{itkh})$ denoting the expected value of outcome Y (i.e., count of low-value services) for beneficiary *i* during year *t* assigned to ACO or non-ACO TIN k living in HRR *h*. "ACO_indicators" is a vector of indicators for each organization in the ACO group, with a single indicator for the control group omitted, "HRR_indicators×Year" is a vector of indicators for each HRR in each year of the sample with a single HRR-year combination omitted, "ACO_contract" is an indicator of being attributed to a Pioneer ACO in 2012, and "Covariates" include patient sociodemographic and clinical covariates

listed above. The β_1 term adjusts for organizations' average level of low-value services in the pre-contract period, and for changes in the distribution of ACO-assigned beneficiaries across ACOs between the pre-contract and post-contract periods. The β_2 term allows for comparison of beneficiaries in the ACO group to control group beneficiaries in the same geographic area, thereby adjusting for region-specific trends in the control group's use of low-value services.

The quantity of interest, β_3, is the mean differential change in low-value services for ACO-attributed beneficiaries relative to local changes in low-value services for the control group. To compare ACOs with higher vs. lower baseline levels of low-value service use, we added to the model interactions between the β_3 term and each of the two measures of ACOs' baseline low-value service use.

A key assumption of this difference-in-differences analysis is that differences in adjusted counts of low-value services between the ACO group and the control group in the pre-contract period would have remained constant in the post-contract period in the absence of the Pioneer program.[15] We tested this assumption by comparing trends in low-value service use between the ACO group and control group over the 2009-2011 pre-contract period (Appendix 2).

We conducted several sensitivity analyses to test for potential sources of bias. First, we adjusted for any differences in trends in low-value service use between the ACO and control groups in the pre-contract period (Appendix 2). Second, we excluded the indicators of service qualification as covariates, in case ACO contracts were associated with changes in the likelihood of patients satisfying qualifying conditions. Third, we tested for differential changes in sociodemographic and clinical characteristics from the pre- to post-contract periods between the ACO and control

groups. If the composition of the ACO and control groups did not change differentially in these observed dimensions, it is less likely that there were differential changes in other, unobserved, dimensions. All analyses employed robust variance estimators clustered at the level of ACOs (for the ACO group) or HRRs (for the control group).[16,17]

2.2.3 RESULTS

The study sample included 18,146,641 person-years (6,110,212 unique beneficiaries), 693,218 in the ACO group and 17,453,423 in the control group. Beneficiary characteristics during the 2009-2011 pre-contract period were similar in the ACO and control groups, adjusted for geographic area, and differential changes in the ACO group were minimal (Table 2.2.2).

During the pre-contract period, the adjusted annual count of low-value services in the ACO group was 1.8 services per 100 beneficiaries lower (P=0.02) than the control group (Table 2.2.3), but trends in the pre-contract period were similar (0.1 services per 100 beneficiaries per year greater for the ACO group; P=0.74). Following the start of Pioneer contracts, there was a differential reduction in the use of low-value services for the ACO group (-0.8 services per 100 beneficiaries; P<0.001), or a reduction of 1.9% relative to the expected 2012 mean for the ACO group of 41.0 services per 100 beneficiaries. Total spending on low-value services in the pre-contract period was similar for the ACO group and control group ($256 per 100 beneficiaries higher in the control group; P=0.13) and trends were also similar ($20 per 100 beneficiaries per year greater for the control group; P=0.88). In 2012, the ACO group underwent a differential reduction in spending of 4.5% (P=0.004).

All clinical categories of low-value services except for preoperative services contributed to the overall differential reduction in the ACO group (Table 2.2.3). The differential reductions were statistically significant for three clinical categories (cancer screening, imaging, and cardiovascular testing and procedures). The greatest absolute reductions in service use occurred for the most frequently delivered services, cancer screening and imaging, which experienced differential reductions of 0.3 services per 100 beneficiaries (P=0.01 and P=0.05, respectively). Cardiovascular testing and procedures underwent the greatest reduction in relative terms, with a differential reduction of 6.3% for the ACO group (P=0.05). Differential reductions in low-value service use were similar in magnitude for higher-priced services (1.4%, 95% CI -0.4%—3.3%) and lower-priced services (2.1%, 95% CI 0.7%—3.5%), as well as for services that were more sensitive to patient preferences (1.7%, 95% CI 0.3%—3.2%) and less sensitive to patient preferences (2.2%, 95% CI 0.7%—3.7%).

ACOs with higher baseline levels of low-value service use than their service area experienced a differential reduction of 1.2 services per 100 beneficiaries (Figure 2.2.1), while ACOs with lower baseline rates experienced a statistically insignificant differential reduction of 0.2 services per 100 beneficiaries (P=0.003 for test of difference in differential reductions between ACO subgroups). Differential reductions in low-value service use were similar for ACOs serving areas with higher or lower baseline levels of low-value service use (P=0.41)

Estimates were not substantially affected by adjusting for small differences in trends in low-value service use during the pre-contract period, or by omitting service qualification indicators from the regression model (Appendix 2).

Table 2.2.2 Beneficiary Characteristics Before and After Start of Pioneer ACO Contracts

	2009-2011		2012			
Characteristic	**Control Group** N=13,041,918	**ACO Group** N=511,426	**Control Group** N=4,411,505	**ACO Group** N=181,792	**Differential Change for ACO Group**	**P-Value**
Age, mean	72.2 ± 0.0	71.9 ± 0.2	72.0 ± 0.0	71.8 ± 0.1	0.1	0.16
Female sex, %	57.4	58.1	57.1	57.8	-0.1	0.54
Race/ethnicity, %						
White	83.2	82.2	82.6	81.7	0.1	0.53
Black	8.6	9.1	8.9	9.2	-0.1	0.13
Hispanic	4.8	5.8	5.0	5.8	-0.2	0.10
Other	3.3	2.9	3.5	3.3	0.2	0.03
Medicaid recipient, %	16.3	16.3	16.1	16.0	-0.1	0.76
Disabled,[a] %	22.0	22.2	22.9	22.8	-0.3	0.06
End-stage renal disease, %	1.2	1.2	1.3	1.2	0.0	0.46
Nursing home resident, %	3.2	2.6	3.1	2.5	-0.1	0.38
CCW conditions[b]						
Total no., mean	5.6 ± 0.0	5.6 ± 0.1	5.8 ± 0.0	5.7 ± 0.1	0.0	0.32
≥6 conditions	47.8	46.9	50.1	48.9	-0.3	0.28
≥9 conditions	19.3	18.6	21.4	20.5	-0.3	0.25
Low-value service measures qualified for,[c] total no., mean	14.9 ± 0.0	14.9 ± 0.0	15.0 ± 0.0	14.9 ± 0.0	0.0	0.47
HCC risk score,[d] mean	1.3 ± 0.0	1.3 ± 0.0	1.3 ± 0.0	1.3 ± 0.0	0.0	0.99
ZCTA-level characteristics, mean						
% below FPL	9.1	8.9	9.1	8.8	0.0	0.83
% with high school degree	75.6	76.3	75.7	76.5	0.1	0.17
% with college degree	19.8	20.6	19.9	20.8	0.1	0.16

ACO = Accountable Care Organization, CCW = Chronic Conditions Warehouse, HCC = Hierarchical Condition Categories, ZCTA = ZIP Code Tabulation Area

Means and percentages were adjusted for geography to reflect comparisons within hospital referral regions. Means are presented ± standard errors.

[a] Refers to beneficiaries for whom disability was the original reason for Medicare eligibility.

Table 2.2.2 (Continued) Beneficiary Characteristics Before and After Start of Pioneer ACO Contracts

[b] Chronic conditions include 25 conditions from the CCW: acute myocardial infarction, Alzheimer's disease, Alzheimer's disease and related disorders or senile dementia, anemia, asthma, atrial fibrillation, benign prostatic hyperplasia, breast cancer, chronic kidney disease, chronic obstructive pulmonary disease, colorectal cancer, depression, diabetes, endometrial cancer, heart failure, hip/pelvic fracture, hyperlipidemia, hypertension, hypothyroidism, ischemic heart disease, lung cancer, osteoporosis, prostate cancer, rheumatoid arthritis/osteoarthritis, stroke/transient ischemic attack.

[c] Refers to the number of low-value service measures that could potentially apply to a beneficiary each year. For example, preoperative testing measures only apply to patients who underwent specific surgical procedures. Qualification criteria for all measures are presented in the eAppendix.

[d] HCC risk scores are calculated based on Medicare enrollment and claims files from the prior calendar year. Higher scores predict higher subsequent spending. Higher scores predict higher subsequent spending.

Table 2.2.3 Differential Changes in Use of Low-Value Services in ACO vs. Control Group

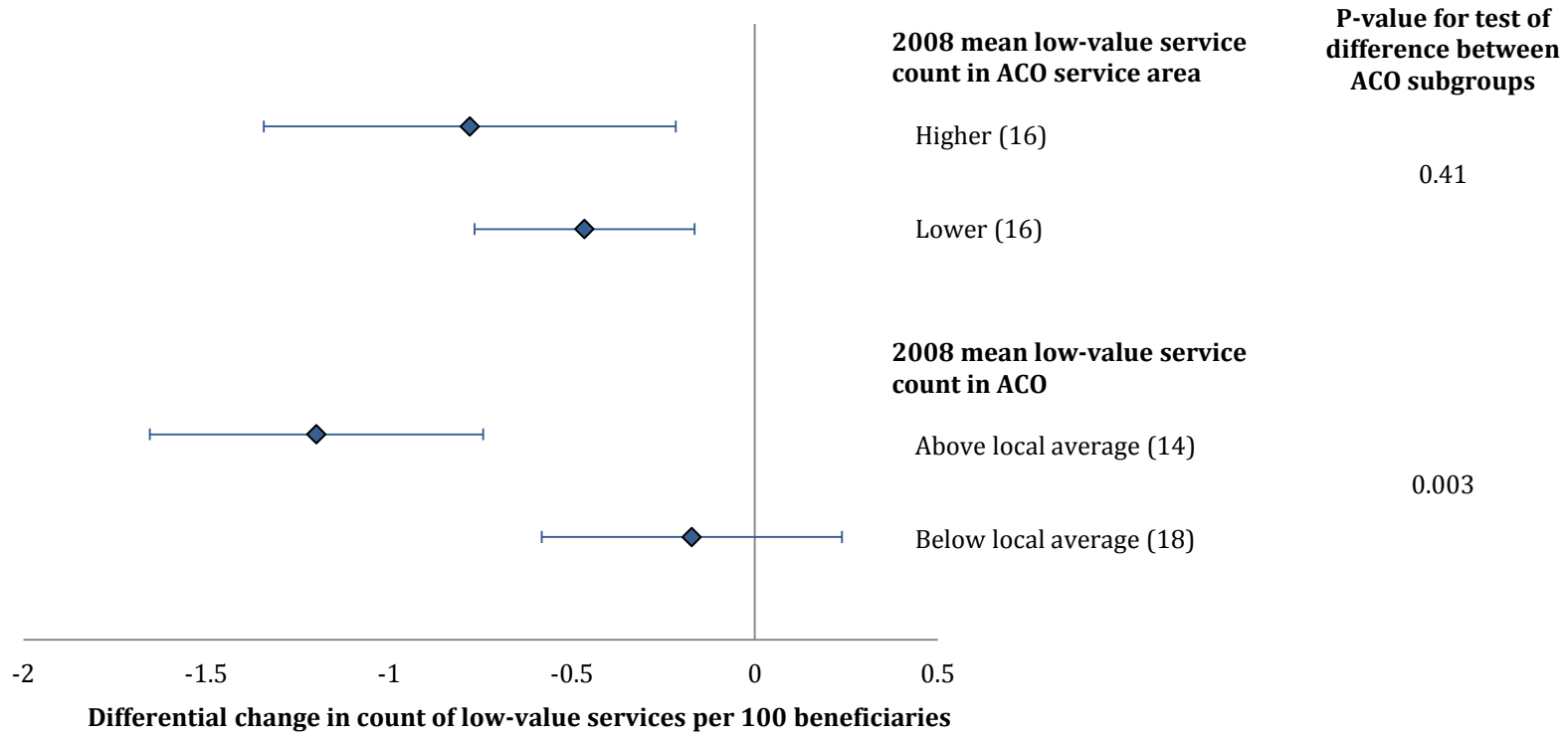
Annual Count or Spending per 100 Beneficiaries	Mean for ACO Group[a]	Baseline Difference between ACO and Control Group	P-Value	Differential Change (per 100 benes)	95% CI	Differential Change as Percent of ACO Mean[b]	95% CI	P-Value
Total low-value services, no.	41.0	-1.8	0.02	-0.8	(-1.2, -0.4)	-1.9	(-2.9, -0.9)	<0.001
Total low-value service spending, $	10301	-256	0.13	-459	(-773, -146)	-4.5	(-7.5, -1.4)	0.004
Low-value services by clinical category, no.[c]								
Cancer screening	11	-0.3	0.27	-0.3	(-0.4, -0.1)	-2.4	(-4.1, -0.7)	0.01
Testing	8.7	-0.7	0.01	-0.2	(-0.5, 0.2)	-1.7	(-5.8, 2.3)	0.39
Preoperative Services	2.1	-0.1	0.01	0.0	(-0.1, 0.1)	1.0	(-3.9, 5.8)	0.69
Imaging	14	-0.6	0.05	-0.3	(-0.5, 0)	-1.8	(-3.6, 0)	0.05
Cardiovascular Tests and Procedures	1.0	0.0	0.43	-0.1	(-0.1, 0)	-6.3	(-12.6, 0)	0.05
Other Invasive Procedures	4.4	-0.1	0.22	-0.1	(-0.2, 0.1)	-1.3	(-4.3, 1.7)	0.38
Low-value services by price, no.[c]								
Higher priced	15	-0.7	0.03	-0.2	(-0.5, 0.1)	-1.4	(-3.3, 0.4)	0.13
Lower priced	25	-1.1	0.03	-0.5	(-0.9, -0.2)	-2.1	(-3.5, -0.7)	0.00
Low-value services by sensitivity to patient preferences, no[c]								
More sensitive	28	-1.4	0.01	-0.5	(-0.9, -0.1)	-1.7	(-3.2, -0.3)	0.02
Less sensitive	13	-0.3	0.17	-0.3	(-0.5, -0.1)	-2.2	(-3.7, -0.7)	0.004

[a] Calculated by as the sum of the 2012 control group mean and the adjusted pre-contract difference between the ACO and control group, which approximates the expected 2012 ACO group mean if there we no differential change.

[b] Calculated as the differential change divided by the mean for ACO group.

[c] Note that the sum of differential changes within each set of service categories equals the total differential change.

FIGURE 2.2.1 Differential Changes in Use of Low-Value Services in ACO vs. Control Group, by Baseline Use



Adjusted differential changes in the total annual count of low-value services for beneficiaries attributed to Pioneer ACOs vs. the control group from the pre-contract period (2009-2011) to the post-contract period (2012) are presented for the following ACO subgroups: (1) ACOs serving areas with a 2008 adjusted count of low-value services per beneficiary in the control group that was greater than vs less than that of the service area of the median ACO, and (2) ACOs with an adjusted count of low-value services per beneficiary in 2008 that was greater vs. less than that of the control group within the ACO's service area. The number of ACOs within each subgroup is indicated parenthetically. Estimates are displayed with 95% confidence intervals and P-values for the difference between subgroups.

2.2.4 Discussion

The first year of the Medicare Pioneer ACO program was associated with a modest reduction in use of low-value services that could be measured directly with claims data. These results are consistent with the hypothesis that global payment initiatives can discourage overuse even while preserving broad provider discretion in determining what services are low-value. Notably, spending on low-value services underwent a differential reduction of 4.5%, substantially larger than the 1.2% overall spending reduction in the first year of the Pioneer program previously estimated with the same methods.[14] This finding suggests that Pioneer ACOs targeted low-value services in their efforts to reduce spending, despite a lack of financial incentives or quality reporting requirements specifically concerning overused services.

Utilization changes appear to have occurred broadly across services, and were not driven by a single measured service or type of service. Even though it may be more difficult for ACOs to incentivize member physicians to reduce higher-priced services, since those services generate more revenue under fee-for-service reimbursement,[18] we observed relative reductions that were similar between higher-priced and lower-priced services. Differential reductions in low-value service use were also similar for services that were more or less sensitive to patient preferences. This finding suggests that reductions in low-value service use in ACOs were driven by changes in physician practice patterns, which accords with research demonstrating that patient preferences are not major obstacles to reducing low-value service use.[19-21]

Reductions in low-value service use were concentrated among ACOs with higher baseline levels of use of these services relative to their service areas, whereas baseline performance of ACO service areas did not predict reductions in low-value service use.

These findings highlight the importance of practice variation within regional markets rather than across markets in predicting organizations' prospects for improving efficiency. In service areas where overuse is especially common, providers may face difficulties in reducing low-value service use markedly below local norms. Bundled payment initiatives may produce greater reductions in overuse if the programs encourage participation of provider organization with more wasteful practices at baseline.

Several limitations of this study warrant discussion. First, organizations selecting into the volunteer Pioneer program may have been uniquely well positioned to identify and reduce wasteful practices. Consequently, similar results may not be achieved if the Pioneer program or similar programs are expanded to include a different set of provider organizations. Second, although our difference-in-differences study design controls for fixed difference between the ACO group and control group, and even though we detected no difference in temporal trends of low-value service use between these groups, it is nevertheless possible that an independent contemporaneous factor affecting ACOs produced a differential change in 2012. It is also possible that organizations entering the Pioneer program may have differentially reduced low-value service use even in the absence of the program. However, we found no evidence that these organizations were experiencing faster reductions in low-value service use prior to the ACO contracts. In addition, reductions in use of higher-priced low-value services would entail a substantial loss in fee-for-service revenue in the absence of ACO contracts, and we found that reductions were unrelated to service price.

Finally, our results do not constitute conclusive evidence of value improvement among Pioneer ACOs. It is possible that important high-value services also experienced

reductions in 2012. Nevertheless, our findings, taken together with studies demonstrating spending reductions greater than Medicare bonus payments[14] and improved or stable performance on measures of patient experiences and quality,[10] are consistent with the conclusion that the overall value of health care provided by Pioneer ACOs improved after their participation in an alternative payment model.

2.2.5 REFERENCES

1. Berwick DM, Hackbarth AD. Eliminating waste in US health care. *JAMA*. 2012;307(14):1513–6.
2. Burwell SM. Setting Value-based payment goals - HHS efforts to improve U.S. health care. *The New England Journal of Medicine*. 2015;(372):897–899.
3. Coulam RF, Gaumer GL. Medicare's prospective payment system: a critical appraisal. *Health Care Financing Review Annual Supplement*. 1991:45–77.
4. McGuire TG, Newhouse JP, Sinaiko AD. An economic history of Medicare part C. *The Milbank Quarterly*. 2011;89(2):289–332.
5. Song Z, Chokshi DA. The role of private payers in payment reform. *JAMA*. 2015;313(1):25–6.
6. Choudhry NK, Rosenthal MB, Milstein A. Assessing the evidence for value-based insurance design. *Health Affairs*. 2010;29(11):1988–94.
7. Elshaug AG, McWilliams JM, Landon BE. The value of low-value lists. *JAMA*. 2013;309(8):775–6.
8. Schwartz AL, Landon BE, Elshaug AG, Chernew ME, McWilliams JM. Measuring low-value care in Medicare. *JAMA Internal Medicine*. 2014;174(7):1067–76.
9. Colla CH. Swimming against the current--what might work to reduce low-value care? *New England Journal of Medicine*. 2014;371(14):1280–3.
10. McWilliams JM, Landon BE, Chernew ME, Zaslavsky AM. Changes in patients' experiences in Medicare Accountable Care Organizations. *New England Journal of Medicine*. 2014;371(18):1715–24.

11. Song Z, Rose S, Safran DG, Landon BE, Day MP, Chernew ME. Changes in health care spending and quality 4 years into global payment. *New England Journal of Medicine*. 2014;371(18):1704–14.

12. Song Z, Safran DG, Landon BE, et al. Health care spending and quality in year 1 of the Alternative Quality Contract. *New England Journal of Medicine*. 2011;365(10):909–18.

13. McWilliams JM, Landon BE, Chernew ME. Changes in health care spending and quality for Medicare beneficiaries associated with a commercial ACO contract. *JAMA*. 2013;310(8):829–36.

14. McWilliams JM, Chernew ME, Landon BE, Schwartz AL. Performance differences in year 1 of Pioneer accountable care organizations. *New England Journal of Medicine*. 2015;in press.

15. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA.* 2014;312(22):2401–2.

16. Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*. 2004;119(1):249–275.

17. Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics*. 2000;56(2):645–6.

18. Landon BE. Keeping score under a global payment system. *New England Journal of Medicine*. 2012;366(5):393–5.

19. Cutler DM, Skinner JS, Stern AD, Wennberg DE. Physician beliefs and patient preferences: a new look at regional variation in health care spending. *National Bureau of Economic Research Working Paper Series*. 2013;19320.

20. Gogineni K, Shuman KL, Chinn D, Gabler NB, Emanuel EJ. Patient Demands and Requests for Cancer Tests and Treatments. *JAMA Oncology*. 2015;Published .
21. Chandra A, Cutler D, Song Z. Who ordered that? The economics of treatment choices in medical care. *Handbook of Health Economics*. 2012;2:397–432.

Chapter 3

Accuracy vs Incentives:
A Tradeoff for Performance Measurement in Health Care and Education

3.1 INTRODUCTION

Perceptions of suboptimal quality in health care and education have spurred interest in promoting performance accountability for hospitals, doctors, schools and teachers. Regulators and institutional purchasers have increasingly employed standardized performance measures for this purpose. In health care, providers are scored on outcomes like mortality, cost, and patient satisfaction, or on processes like rates of appropriately prescribing a medication. In education, there is substantial interest in value-added modeling, which assesses a teacher or school's performance based on changes in student test scores. Several federal and state policies in the United States have accelerated these trends, mandating public disclosure of certain performance measures and tying substantial financial incentives to others.

The reliability of performance measures in these settings has been a persistent concern (Hofer et al. 1999, Kane and Staiger 2002). High variance of measured outcomes and relatively small sample sizes of patients or students can result in substantial measurement error. Outstanding performers in one period often experience reversion to the mean soon after, suggesting that initial performance was partially due to chance. To address this limitation, it is common to modify estimates of observed performance by shrinking them toward a common prior value, typically the average observed performance of all agents. A substantial literature dating to Stein (1956) illustrates that shrinkage estimation reduces measurement error, and shrinkage estimation is employed in a variety of specific modeling strategies referred to as mixed, hierarchical, multilevel or random effects modeling, or empirical Bayes estimation. Research on the policy applications of these techniques has focused on their statistical properties like precision or bias (Koedel, Mihaly, and Rockoff 2015; Normand and

Shahian 2007). However, because these measures are intended to affect the market behavior of consumers and suppliers, these measures should be ultimately judged on the basis of economic rather than statistical criteria.

This paper explores the implications of applying Bayesian shrinkage techniques to performance measurement in public policy. The study is motivated by a simple observation: shrinkage estimation reduces a measure's responsiveness to measured behavior. When shrinkage techniques are not employed, and performance is estimated as the mean of an agent's observed performance (i.e. the average mortality of a surgeon's patients), then an increase in that agent's true performance will coincide with an equal expected increase in measured performance. Adjusting these estimates using shrinkage techniques will tend to increase the measured performance of below-average agents, and decrease the measured performance of above-average agents. In both cases, however, the shrinkage estimate will be less responsive to the agent's observed performance and to the agent's true unobserved performance. For incentive schemes in which agents are rewarded according to their measured performance, reducing the responsiveness of a measure will reduce the marginal incentive for performance improvement. Thus, the incentive properties of performance estimation techniques, which are economic properties, are a first-order concern for designing optimal incentive schemes in public policy.

This observation motivates several policy-relevant questions. Do measure accuracy and measure responsiveness both contribute to the success of accountability-based public policies? If so, in what cases does one property contribute more to welfare than the other? In the context of current and potential policies, is the loss in measure responsiveness from shrinkage estimation substantial compared to the accuracy gains?

Is the tradeoff avoidable? This paper aims to provide theoretical and empirical traction on these unexplored issues.

The paper's first contribution is an assessment of the welfare implications of accuracy and responsiveness in performance estimation. In a stylized model of two agents with market power competing on quality, accuracy and responsiveness of performance signals each contribute to welfare. Greater accuracy improves consumers' match to agents, thereby reducing welfare losses resulting from misinformed consumer choices. Responsive performance estimation drives a demand response to quality, reducing welfare loss arising from suboptimal quality investment by agents. The relative welfare contribution of each measurement property depends on the policy setting. This model motivates my examination of shrinkage estimation by demonstrating that accuracy of performance measurement and incentives for performance improvement can be substitutes in promoting welfare.

My second contribution is a characterization of the tradeoff between accuracy improvement and reduced measure responsiveness entailed by shrinkage estimation. Shrinkage estimators reduce measure responsiveness by one minus the shrinkage factor, which depends on sample size, variance in true performance across agents, and variance in observed performance within agents (i.e. noise). Thus, incentives are distorted by the extent to which an agent's performance score is determined by other agents' observed performance, resulting in a free-riding problem. Greater incentive distortions are to be expected whenever observed performance is an especially noisy signal of true performance. In particular, agents with few available performance observations, like teachers with small classrooms or hospitals with few patients, face

weaker incentives to improve performance. I also demonstrate the connection between the accuracy-incentives tradeoff and alternate definitions of a measure's biasedness.

Third, I show that the magnitude of the accuracy-incentives tradeoff is substantial in the context of a hospital performance measurement. Using a Monte Carlo simulation, I examine accuracy and responsiveness in the context of measuring heart attack mortality, which is publicly reported in a national disclosure program. I calculate that the current preferred method for shrinking one-year performance estimates reduces measure responsiveness by 34 percent on average, and by 64 percent for smaller hospitals. These smaller hospitals must decrease mortality by 2.8 times more than a large hospital in order to experience an equal measured mortality improvement.

Finally, I compare the accuracy and responsiveness of several alternate approaches to estimating hospital performance. Although shrinkage estimators tend to reduces measurement error substantially, similar reductions in error can be achieved without shrinkage by increasing the number of years used to estimate performance. Scoring each estimation technique based on accuracy and responsiveness, I identify a frontier of techniques that dominate others. Notably, the current risk-standardized mortality rate measurement employed by Medicare is dominated.

This study is related most closely to the economics literature on performance measurement in health care and education. In education, studies have highlighted the obstacles introduced by imprecise performance measures (Kane and Staiger 2002; Staiger and Rockoff 2010). A review of value-added modeling documents many studies of measurement properties like bias and stability (Koedel, Mihaly, and Rockoff 2015). Health economics is largely devoted to understanding extensive information imperfections in health care (Arrow 1963) which may motivate quality reporting or

pay-for-performance schemes (Kolstad 2013; Richardson 2013). Many of these studies of health care and education fall within a broader economic literature on quality disclosure and certification (Dranove and Jin 2010). There is also an expansive statistical, medical, and policy literature on various properties of health care quality measurement, including measure reliability (e.g. Adams et al., 2010; Dimick et al., 2004; Nyweide et al., 2009).

A broader economics literature concerns the consequences of imprecise quality signals in markets. In organizational economics, this topic has been a particular area of focus, especially with regard to the optimal power of incentive contracts (Gibbons and Roberts 2013). Precision of performance signals also plays a role in the economics of discrimination. For example, the canonical Phelps (1972) study of statistical discrimination concludes by illustrating how high-performing minorities may face discrimination in the labor market if they produce a high variance performance signal. Just as this penalty for high-variance performance may reduce human capital investment (Farmer and Terrell 1996), I argue that shrinkage estimation in education and health care may discourage quality investment. The key distinction between my research and the broader literature on quality signals is my focus on the public policy setting. For example, I do not impose labor market equilibrium conditions equating compensation to workers' expected productivity. Instead, I assume that a government paying for health care or education services differs from other employers in that the government can provide compensation that departs from a posterior belief about workers' productivity. This ensures that whether to use shrinkage estimation for performance measures is the government's choice rather than a necessity.

The remainder of the paper proceeds as follows. Section 3.2 presents a stylized model demonstrating the contributions of measure accuracy and performance incentives to welfare. Section 3.3 describes the accuracy-incentives tradeoff in the context of shrinkage estimation. Section 3.4 details the Monte Carlo simulation and its results. Section 3.5 discusses several implications of the analysis for health and education policy and provides a brief conclusion.

3.2 Measurement Accuracy, Measurement Responsiveness, and Welfare: A Stylized Model

I consider a stylized model in which agents with market power choose levels of quality and quality signals guide consumers' choice of agents. This model could describe patients choosing among medical provider or students choosing schools. In equilibrium, I find that total welfare is a function of the magnitude of signal error and the responsiveness of the signal to agents' quality choices. The intuition for this result follows from the two ways in which quality information contributes to welfare. First, accurate quality signals promote efficient sorting for consumers, who might choose an inferior agent based on an erroneous quality signal. Second, quality signals that are responsive to agent behavior elicit a demand response, increasing agents' incentives for investing in quality and raising quality above suboptimal levels. I also highlight two special cases which illustrate the context-dependence of whether signal accuracy or performance incentives are more important contributors to welfare. In the first, welfare gains can only be achieved through improved signal accuracy because quality levels are exogenous. In the second, welfare gains can only be achieved through increased financial incentives for quality because demand is unresponsive to quality.

Consumers are arrayed uniformly on a line between two agents (agent A and agent B), with $z \in (0,1)$ denoting a consumer's distance from agent A. Both the distance between agents and the number of consumers are normalized to one. The model proceeds in three stages. First, each agent *j* simultaneously chooses a level of quality $u_j \in [0, \infty)$ and bears the costs of that quality investment. Second, consumers perceive a quality signal from each agent. Third, consumers sort between agents, who receive a regulated fee for each consumer they serve.

Consumer utility depends on quality of the consumer's chosen agent, u_a or u_b, and transport costs $c > 0$ per unit of travel. Specifically,

$$U(z) = \begin{cases} \alpha + u_A - cz & \text{if } j = A \\ \alpha + u_B - c(1-z) & \text{if } j = B \end{cases}$$

I assume $\alpha > c/2$, which ensures that the minimal utility achieved from being served by an agent exceeds the maximum transport costs entailed by choosing the closest agent. Thus, each consumer will choose one of the agents. Consumers perceive agent quality as $\hat{u}_j$, a signal of quality that contains some error ε_j. By definition:

$$\hat{u}_j = u_j + \varepsilon_j$$

Quality signals are perceived identically by all consumers, and no particular distribution of the error is assumed. I do assume that $E\left[\varepsilon_j \middle| u_{-j}, \varepsilon_{-j}\right] = E\left[\varepsilon_j\right]$, where *–j* indicates the agent who is not agent *j.* Thus, error in one agent's quality signal is unaffected by the other agent's true quality or the error in their quality. Note that an agent's quality signal and the consumer's perception of that agent's quality are equal. This is reasonable in the public policy settings I discuss, where information imperfections are common, and consumers may not have access to multiple sources of reliable performance

information. The difference in the quality signals from each agent yields a relative quality signal, which is represented by the following notation:

$$\hat{u}_j^{\Delta} = u_j - u_{-j} + \varepsilon_j - \varepsilon_{-j} = u_j^{\Delta} + \varepsilon_j^{\Delta}$$

These quality signals yield a demand for each agent, $Z_j(\hat{u}_j^{\Delta})$.

Agents are self-interested, and their utility is the difference between revenue and effort costs, which I assume to be quadratic. Agents receive a regulated price r for each consumer they serve. Thus, physician utility is

$$V_j = Z_j r - \frac{1}{2}u_j^2$$

Specifying effort costs as quadratic function of quality conveniently ensures an interior solution for agent choice of quality. Revenue and costs are the same for both agents, implying symmetric behavior in equilibrium. Note also that agents are risk-neutral.[1]

Equilibrium

Consumers maximize utility on the basis of the perceived quality of both agents, choosing agent A if and only if $\hat{u}_A^{\Delta} > cz - c(1-z)$, yielding the following demand:

$$Z_j = \frac{1}{2} + \frac{\hat{u}_i^{\Delta}}{2c}$$

To ensure an interior sorting solution, I assume that $\max(|\hat{u}_j^{\Delta}|) < c$. In a symmetric equilibrium with equal agent quality, this corresponds to the assumption that $\max(|\varepsilon_j^{\Delta}|) < c$.

[1] At this point, it bears emphasizing the stylized nature of this model. In order to consider issues of performance signal accuracy and responsiveness in isolation, the model does not incorporate additional concerns regarding agent altruism (Kolstad 2013, McGuire 2000), multitasking (Holmstrom and Milgrom 1990), the insurance value of performance contracts to agents (Gibbons and Roberts 2013), or agent participation decisions (Rothstein 2015).

Providers maximize their utility, with the following first order condition describing their choice of quality:

$$u_j = \frac{dE[Z_j]}{du_j} r$$

Substituting for the derivative of demand and noting that $\frac{dE[\hat{u}_j^{\Delta}]}{du_j} = \frac{dE[\hat{u}_j]}{du_j}$ (which follows from the prior assumption that $E[\varepsilon_j | u_{-j}, \varepsilon_{-j}] = E[\varepsilon_j]$, yields equilibrium quality supply:

$$u_j = \frac{dE[\hat{u}_j]}{du_j} \frac{r}{2c}$$

Given that the right hand side terms of this expression are equal for both agents, agent quality choices are indeed identical, which means that equilibrium demand is

$$Z_j = \frac{1}{2} + \frac{\varepsilon_j^{\Delta}}{2c}$$

Welfare

Before characterizing welfare in equilibrium, it is instructive to consider the welfare resulting from various potential sorting and quality decisions. If agent quality choice is symmetric and consumers choose agents such that all consumers located at $z < Z_A$ choose agent A and all located at $z > Z_A$ choose agent B, then realized total welfare following sorting can be expressed as

$$\int_0^{Z_A} (\alpha + u - cz)\, dz + \int_{Z_A}^1 \big(\alpha + u - c(1-z)\big) dz - u^2$$

Solving and rearranging yields

$$\alpha + \frac{1-c}{4} - \left(u - \frac{1}{2}\right)^2 - c\left(Z_A - \frac{1}{2}\right)^2$$

The leftmost two terms, consisting of constants, reflect the maximum total realized utility for agents and consumers. This first-best utility can only be achieved when quality equals 0.5 and demand equals 0.5 for each agent. For intuition behind this result, note that optimal quality entails equalizing the marginal cost of quality for both agents, $2u$, with the marginal benefit for consumers, u. Because agents choose equal quality in equilibrium, it follows that optimal sorting occurs at the midpoint between agents, which minimizes travel distance.

The expected total welfare across a range of possible error draws is simply the expectation of this expression. Evaluating the expectation and substituting demand and quality in equilibrium and yields the following expression for expected welfare loss in equilibrium relative to the first-best scenario:

$$\left(\frac{dE[\hat{u}_j]}{du_j}\frac{r}{2c} - \frac{1}{2}\right)^2 + \frac{1}{4c}E\left[\left(\hat{u}_j^{\Delta} - u_j^{\Delta}\right)^2\right]$$

Note that $E\left[\left(\hat{u}_j^{\Delta} - u_j^{\Delta}\right)^2\right]$ is the mean squared error of $\hat{u}_j^{\Delta}$, the relative quality signal, as an estimate of u_j^{Δ}, the true difference in agent quality. The two terms in this expression represent two components of welfare loss in equilibrium. The left term is the square of the difference between equilibrium quality and optimal quality. If perceived signals of agent quality are fully responsive to agent quality investments ($\frac{dE[\hat{u}_j]}{du_j} = 1$), then quality investment will be suboptimal when $r < c$. Welfare losses from suboptimal quality will be exacerbated when quality signals are less than fully responsive ($\frac{dE[\hat{u}_j]}{du_j} < 1$). The right term represents the welfare loss attributable to excess travel costs due to error in the relative quality signal. When quality signals have greater error, a greater number of consumers near the margin of agent selection will make inefficient agent choices. I say

an *accuracy-incentives tradeoff* exists when one component of this welfare expression increases while the other component decreases.

Policy Responses and Special Cases

These equilibrium conditions suggest that quality disclosure policies and performance payment policies may be welfare-improving. A regulator can promote optimal sorting by reducing the mean squared error of quality signals that consumers perceive. For example, if the government were able to measure and publicly disclose a quality signal with zero error, then optimal sorting would result. Because mean squared error is the sum of variance and squared bias, note that an unbiased signal with large variance may produce greater welfare loss than a biased signal with lesser variance. A regulator can induce optimal agent effort in two ways. The regulated fee can be set so that $r = c\left(\frac{dE[\hat{u}_j]}{du_j}\right)^{-1}$. However, this approach would entail extremely high fees in the event of low demand elasticity (i.e. high transport costs). Alternatively, the regulator can introduce a bonus payment for quality such that agents now receive payments of $Z_j r + b\hat{u}_j$. Agents would now choose quality levels on the basis of the following first order condition

$$\frac{dE[\hat{u}_j]}{du_j}\frac{r}{2c} + \frac{dE[\hat{u}_j]}{du_j}b = u_j$$

Optimal effort can be achieved by choosing a bonus payment *b* and regulated fee *r* such that the left hand side equals 0.5, the optimal choice.

Although both the provision of accurate performance information and appropriate quality incentives can be welfare-improving, the relative welfare contribution of these policies will depend on the setting. I now illustrate two special

cases in which only one of these policy mechanisms contributes to welfare. First, consider a setting in which agent quality is fixed and depends only on innate talent, rather than effort. In this case, u_i is no longer a choice variable for agents. In this setting, demand equals $u_j^{\Delta} + \frac{\varepsilon_j^{\Delta}}{2c}$ and welfare losses relative to the first best equal $\frac{1}{4c} E\left[\left(\hat{u}_j^{\Delta} - u_j^{\Delta}\right)^2\right]$. First-best welfare can be achieved by providing quality signals with zero error. However, policies that provide incentives for quality have no effect on welfare, since quality is exogenous. Second, consider a setting in which demand is inelastic with respect to quality. In the setting of this stylized model, this assumption can be modeled as transport costs c approaching infinity. In this setting, T_i equals one half and consumers are split equally between agents regardless of quality differences. In the absence of a quality bonus payment scheme, equilibrium quality is zero, and relative to the first best, welfare loss equals $\left(\frac{dE[\hat{u}_j]}{du_j} b - \frac{1}{2}\right)$. First-best welfare can be achieved by providing an appropriate bonus for measured quality. However, improving measure accuracy has no effect on welfare.

To summarize, in this stylized principal-agent model, welfare loss can be decomposed into two sources: consumer uncertainty about quality and insufficient (or excessive) incentives for quality improvement. This welfare decomposition and the subsequent discussion of optimal policy responses illustrate how these two sources of welfare loss are addressed separately by policies like disclosure of precise performance information or provision of performance bonuses. Whether improving accuracy of performance signals or improving incentives for performance improvement contributes more to welfare improvement depends on demand responsiveness and the marginal

costs of quality improvement. However, there will be an accuracy-incentive tradeoff when one component of welfare loss increases and the other decreases.

3.3. Shrinkage Estimation and the Accuracy-Incentives Tradeoff

Shrinkage estimation describes a broad class of estimation techniques that adjust raw observed estimates toward a common prior value. These estimates are said to "borrow strength" or "borrow information" across units of observation, because the parameter of one unit is estimated using data from an independent unit. In the context of performance estimation, this means that estimates of an agent's performance will depend on other agents' performance. Early motivation for such approaches was provided by Stein (1956), who proved the paradoxical result that, when estimating the means of several independent normal random variables, simple averages were inferior with respect to mean squared error to an alternative estimation approaches. The massive breadth of the ensuing literature precludes a comprehensive review here. In this section, I briefly review general properties of shrinkage estimators and demonstrate how choosing between estimators with and without shrinkage can entail an accuracy-incentive tradeoff.

Consider estimating the performance of many educators or health care providers. Assume a data-generating process for the health or educational outcomes of individuals *i* who receive services from one of agents *j*:

$$y_{ij} = \beta x_{ij} + u_j + \varepsilon_{ij}, \qquad i = 1, \dots, n_j$$

where x is a vector of individual covariates, u_j is the agent performance and ε_{ij} is error. Agent performance is assumed to be independently normally distributed with mean μ_u.

A standard shrinkage estimator for u_j, $\tilde{u}_j$ can be expressed as follows: (Gelman and Hill 2007; Koedel, Mihaly and Rockoff 2015, Skrondal and Rabe-Hesketh 2009):

$$\tilde{u}_j = s_j \bar{r}_j + \left(1 - s_j\right) \widehat{\mu_u}$$

where $\bar{r}_j$ is agent's average residuals $\bar{y}_j - \hat{\beta}\bar{x}_j$, and $s_j \in [0,1]$ is the shrinkage factor, which equals $\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_\varepsilon^2}{n_j}}$, where σ_u^2 and σ_ε^2 are variances of u_j and ε_{ij} (i.e. the across-agent and within-agent variance).[2] If the model is correctly specified, then the shrinkage estimator minimizes the estimates' mean of the squared errors, and is particularly useful in correcting for attenuation bias when performance estimates are used as regressors (e.g. Chandra et al. 2013, Chetty, Friedman, and Rockoff 2014).

Several variants of such shrinkage estimation have been employed for health care and education measurement. Because larger hospitals often exhibit performance superior to smaller hospitals due to scale economies or learning-by-doing (Gaynor, Seider, and Vogt 2005), alternative methods shrink observed hospital outcomes toward a volume-standardized performance mean rather than an overall mean (Dimick et al. 2009, Silber et al. 2010). Shrinkage estimates of teacher value-added can decompose the source of within-teacher variance into idiosyncratic annual classroom effects (i.e. exogenous classroom shocks) and student effects (Kane and Staiger, 2008). Others account for drift in teacher quality over time (Chetty, Friedman, and Rockoff 2014). Despite these differences, the methods all attempt to reduce mean squared error of

[2] Note that the estimate $\tilde{u}_j$ cannot be operationalized as written since the equation requires estimates of the variance components and mean. This is a general property of shrinkage estimators, and there are many alternate ways of incorporating estimates of these parameters into the calculation of $\hat{u}_j$. Fully Bayesian approaches employ a posterior distribution of these additional parameters (estimated based on prior distributions) while empirical Bayes approaches plug in point estimates. For details and examples, see Gelman and Hill (2009), Gelman et al. (2014), Morris (1983), Guarino et al. (2014), Chetty, Friedman, and Rockoff (2014), Dimick, Staiger, and Birkmeyer (2010).

performance estimates by shrinking an average residual toward a common value, with the magnitude of shrinkage depending on a decomposition of variance.

To illustrate several properties of shrinkage estimation graphically, I briefly compare observed and shrunken performance estimates ($\bar{r}_j$ and $\tilde{u}_j$) obtained from synthetic classroom data for 100,000 teachers. In these data, teachers serve classrooms of 27 students, true teacher performance u_j is independently distributed$N(0,0.15)$, and a student outcome is independently distributed $N(u_j, 0.95)$. This produces a shrinkage factor of 0.4 for each teacher performance estimate. These parameters were chosen based on the distribution of teacher ability and the measure reliability found in Rothstein (2015). In this example, since no student covariates must be adjusted for, the mean of student outcomes within a class constitutes a teacher's observed performance, $\bar{r}_j$. Shrunken posterior performance estimates, $\tilde{u}_j$, are obtained via multilevel modeling with a random teacher effect. Panel A of Figure 3.1 shows the distribution of observed and shrunken performance estimates. Because of within-classroom variance in student performance, observed performance is over-dispersed relative to true teacher ability. Shrunken performance estimates exhibit less variance than both observed and true teacher performance. As noted by Chandra et al. (2013), the latter property follows because true performance is the sum of the shrunken performance prediction and an orthogonal prediction error.

Shrinkage estimators may be considered unbiased or biased depending on the criteria for bias. Consistent with this observation are Panels B and C of Figure 3.1, which present binned scatterplots of true performance vs measured performance and vice versa. First, consider for some performance measure $\hat{u}_j$, the property $E\left[u_j|\hat{u}_j\right] = \hat{u}_j$, which I refer to as prediction unbiasedness, following Chetty Friedman and Rockoff

(2014). If a measure is prediction unbiased, then an agent's measured performance will equal his or her expected true performance. As shown in Panel B, these shrunken performance measures very closely approximate the conditional mean of true teacher performance in the synthetic data. For this reason, linear shrinkage estimators are sometimes referred to as best linear unbiased predictors (BLUPs) (Skrondal and Rabe-Hesketh 2009).[3] Alternatively, observed performance clearly demonstrates prediction biasedness, overestimating the performance of teachers with relatively greater observed performance and underestimating the performance of teachers with inferior performance. Second, consider measurement unbiasedness, defined here as $E\left[\hat{u}_j | u_j\right] = \hat{u}_j$. If a measure exhibits measurement unbiasedness, then an agent's expected measured performance will equal their true performance. As shown in Panel C, shrinkage estimators do poorly according to this criteria, underestimating the performance of high performers and overestimating the performance of low performers. Alternatively, true performance very closely approximates the conditional average of observed (unshrunken) teacher performance.

Measurement bias relates to a key incentive property of shrinkage estimators: responsiveness to agent behavior. I define measure responsiveness as $\frac{dE[\hat{u}_j]}{du_j}$, the change in expected measured performance for a change in actual performance. Assuming that $E[\,\varepsilon_{ij} \big|\, u_j] = 0$ and that an individual agent's performance contributes negligibly to the average performance, $\frac{dE[\hat{u}_j]}{du_j} = \frac{dE[\tilde{u}_j]}{du_j} \approx s_j$. Thus, the size of the shrinkage factor faced by an agent equals the measure's responsiveness to that agent's behavior. (Note that in

[3] There has been recent interest in questioning whether assumptions required for unbiasedness hold when consumers' choice of agents is not exogenous. See, for example Kalbfleisch and Wolfe (2013) and Guarino et al. (2014).

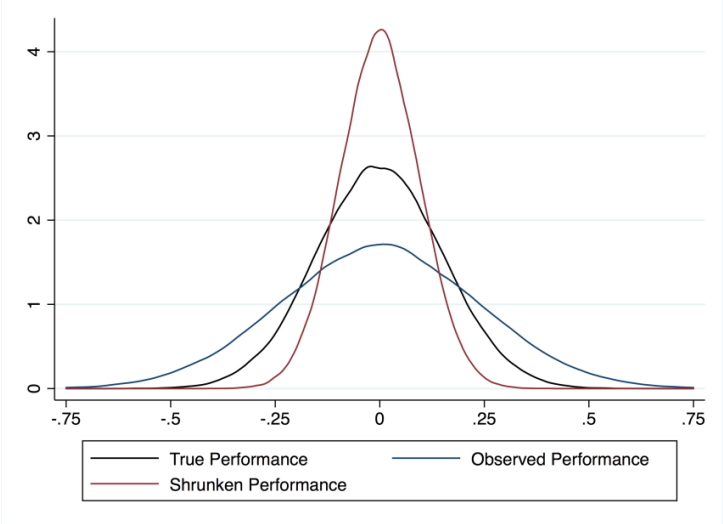
Panel C of Figure 3.1, the scatter plot slope for shrunken estimates indeed equals the shrinkage factor 0.4.) The loss of measure responsiveness entailed by shrinkage estimation equals one minus the shrinkage factor. As the shrinkage factor approaches one, the measure becomes fully responsive to agent behavior, since $\frac{dE[\bar{r}_j]}{du_j} = 1$. In this case, the group mean no longer contributes to the shrinkage estimate, which converges to its fixed effect estimate. Because measure responsiveness is increasing in n_j and $\hat{\sigma}_u^2$ and decreasing in $\hat{\sigma}_\varepsilon^2$, shrinkage estimates of performance will be less responsive for agents serving a smaller number of consumers (e.g. smaller hospitals), for measured outcomes with a large amount of residual error, and for settings in which agent performance is very similar.

Consider these shrinkage estimation properties in light of the model presented in Section 3.2. Recall that equilibrium welfare loss equals $\left(\frac{dE[\hat{u}_j]}{du_j}\frac{r}{2c} - \frac{1}{2}\right)^2 + \frac{1}{4c}MSE\left[\hat{u}_j^{\Delta}\right]$. Welfare loss is increasing in the mean square error of the relative performance signal and, when $r < c$, decreasing in the responsiveness quality signals to true quality. Because shrinkage estimation decreases mean squared error but reduces $\frac{dE[\hat{u}_j]}{du_j}$ from 1 to s_j, adopting shrinkage estimation entails an accuracy-incentives tradeoff. For the intuition behind this result, recall that agents' investments in quality increase with the responsiveness of quality signals to behavior. For example, the quality bonus scheme described in Section 3.2 consists of a linear schedule of reward payments for measured performance. As measure responsiveness decreases, so does the effective marginal quality bonus for improved performance. Consumers' demand response to quality, another driver of agent quality investments in Section 3.2, will be similarly diluted if publicly disclosed quality ratings are shrunken and consumers perceive these ratings to

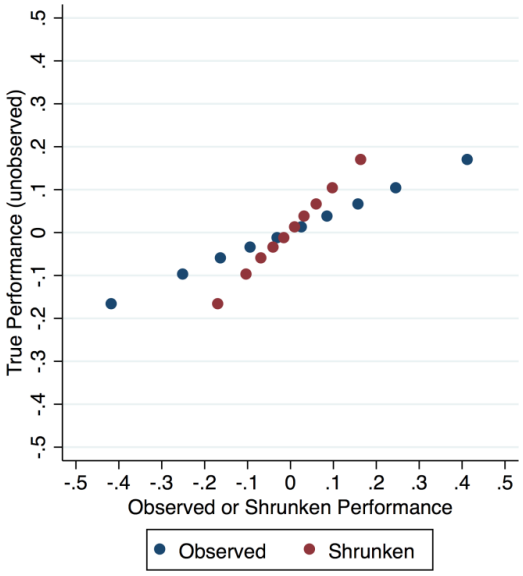
be measurement unbiased. Thus, although shrinkage estimators may improve consumer sorting by improving measure accuracy, they reduce measure responsiveness and may lead to insufficient quality investment.

Figure 3.1 Bias in Observed and Shrunken Performance

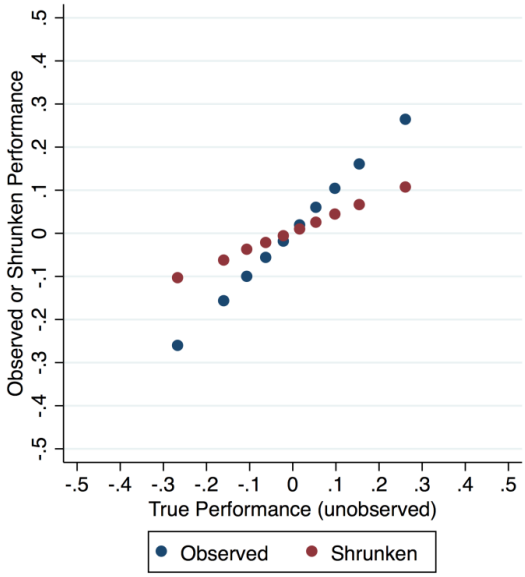
A) Distribution of True and Measured Performance



B) Prediction Bias of Observed and Shrunken Performance



C) Measurement Bias of Observed and Shrunken Performance



Notes: These figures compare true performance (unobservable) to observed performance and shrunken estimates of observed performance. The data are simulated to match the variance properties of teacher value-added measures. True performance is distributed normally with mean zero and σ_u of 0.15. With classrooms of 27 students and σ_e of 0.95, reliability is 0.4 and the shrinkage weight is 0.6. Observed performance is the average outcome within in a teacher's classroom. Shrunken performance is predicted via random effects estimation. Panel A is a kernel density plot of true performance, observed performance, and shrunken performance estimates for 100,000 teachers. Panels B and C present binned scatterplots, which were constructed by dividing teachers into deciles based on their horizontal axis values and plotting means within each decile.

3.4. Hospital Quality Measurement and the Magnitude of the Accuracy-Incentives Tradeoff

I use simulation to assess the magnitude of the accuracy-incentives tradeoff in the case of hospital performance measurement. Currently, the Centers for Medicare and Medicaid Services (CMS) employs shrinkage estimation to evaluate hospital mortality rates and rates of readmissions for patients with select diagnoses. These measures, constructed from Medicare claims data, are part of broader efforts to tie Medicare payments to measures of health care value (Burwell 2015) and to report hospital quality ratings (Werner and Bradlow 2006). 30-day readmission rates have been publicly reported since 2009 and began contributing to hospital payment penalties through the Hospital Readmissions Reduction Program in fiscal year 2013. 30-day mortality ratings have been publicly reported since 2007 and began contributing to hospital payment adjustments as part of the Medicare Hospital Value-Based Purchasing Program in the 2014 fiscal year.[4] CMS methods for calculating mortality and readmissions measures are broadly similar, involving hierarchical logistic models that include patient characteristics as covariates (Ash et al., 2012; Krumholz et al., 2006).

Simulation Methods

I use Monte Carlo simulation to study measurement of hospital 30-day mortality for patients with acute myocardial infarction (AMI), also known as heart attack. AMI

[4] Hospital Readmissions Reduction Program penalties take the form of reductions in Medicare payments for all hospital admissions. The reduction is based on a hospital's risk-adjusted readmissions rates for patients admitted with a select set of diagnoses. A hospital's penalty is equal to the proportion of Medicare payments for these admissions that can attributed to readmissions in excess of a hospital's expected number of readmissions, with a maximum penalty is a 3% in fiscal year 2015. Payments for the Hospital Value-Based Purchasing Program payments are more complex. In fiscal year 2015, 1.5% of base hospital payments were withheld from participating hospitals, and this money was used for incentive payments. Payments were calculated on the basis of 26 performance measures, which are combined into composite scores for achievement as well as improvement.

was one of the first diagnoses used for CMS mortality measures, and is a serious complication of cardiovascular disease, the leading cause of death in the United States. The simulation is compares the performance of alternative measurement techniques according to two properties: root mean squared error (accuracy) and measure responsiveness (incentives). The simulation allows me to construct true hospital performance, which is typically unobserved, and to calculate an error equal to the difference between this value and measured performance. In addition, by taking repeated draws of data, simulation results incorporate findings from a broad set of possible hospital outcomes. Many studies use simulation to examine the properties of performance measures in health and education (Normand et al., 2007; Thomas and Hofer, 1999; Koedel Mihaly and Rockoff 2015; Rothstein 2015). My analysis closely follows that of Ryan et al. (2012), which compared the accuracy of several alternate AMI mortality measures. I replicate and extend those simulation methods by assessing measure responsiveness in addition to measurement error. The simulation methods, briefly described here, are detailed more fully in Ryan et al. (2012).

The data generating process has been calibrated to approximate the distribution of risk-adjusted mortality in Medicare inpatient claims data. In addition, the simulation includes a rejection sampling condition that discards any simulation iteration in which the simulated data differ substantially from Medicare inpatient data in more than one of several moment conditions.[5] These conditions, and their values in Medicare inpatient data are: mean mortality (0.209), within-hospital standard deviation in mortality (0.091), between-hospital standard deviation in mortality (0.078), mean annual change in mortality (-0.007), within-hospital standard deviation of annual mortality change

[5] Specifically, this meant that the iteration was discarded if more than one of the simulated data parameters fell outside of a bootstrapped 95% confidence interval of the Medicare data parameter.

(0.137), between-hospital standard deviation of annual mortality changes (0.031), and mean hospital AMI volume (104.8).

The data generation process involves the following steps. For each of 3000 hospitals, an initial volume of AMI patients and an annual growth rate in volume are drawn from a truncated gamma distribution and a normal distribution, respectively (see Ryan et al. [2012] for all parameter values). Each hospital is assigned an initial raw mortality rate and an annual growth rate in mortality improvement, drawn from normal distributions. Annual raw mortality rates are then adjusted to reflect improved mortality in higher volume hospitals. Specifically, raw mortality rates are adjusted based on annual hospital volume and the empirical relationship between volume and risk-adjusted mortality in Medicare inpatient claims, which was modeled using a generalized linear model (Bernoulli family, logit link) and a of 5^{th} degree polynomial function of hospital volume. The resulting annual mortality rate serves as a hospital's true mortality score and corresponds to each patient's probability of dying within 30 days of admission. Deaths are assigned according to a random draw for each patient. Note that the probability of mortality is not a function of patient characteristics. This corresponds to an assumption that risk-adjustment eliminates residual confounding in all mortality measurement techniques I consider.

For each measurement technique that I consider, I calculate hospital mortality scores based on one, two, or three years of observed mortality. In each simulation iteration, the accuracy of each measure is assessed by comparing measured mortality scores $\hat{u}_j$ to true mortality in the following year u_j. The temporal lag reflects the role of public reporting policies in providing past hospital performance data to inform current patient decisions. Measure accuracy is scored as root mean square error (RMSE),

$\sqrt{\overline{(\hat{u}_j - u_j)^2}}$. Each measure's responsiveness, $\frac{dE[\hat{u}_j]}{du_j}$ is scored as the average shrinkage weight s_j. Accuracy and responsiveness are assessed across all hospitals and separately for hospitals categorized as small, medium and large, where patient volume for small hospitals is below the 25th percentile (approximately 30 AMI admissions) and volume for large hospitals is above the 75th percentile (approximately 143 AMI admissions).

I consider five alternate measures of hospital mortality, four of which are included in Ryan et al. (2012). The first measure is observed over expected mortality (OE). OE, which is not a shrinkage estimator, has been used to estimate cardiac surgery performance (Kolstad 2013). It is calculated as follows:

$$\widehat{OE}_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{\sum_{i=1}^{n_j} \hat{\beta}_0 + \hat{\beta}_1 X_{ij}} \cdot \bar{y}$$

where y_{ij} is an indicator for death, $\bar{y}$ is the overall average mortality rate, and X is a vector of patient characteristics. The denominator is the expected number of patient deaths based on prediction via linear regression. In the absence of patient covariates, this expression simplifies to the observed mortality rate $\sum_{i=1}^{n_j} y_{ij} / n_j$. I also implement a moving average (MA) of this estimator, a simple average of OE estimates over two or three years. Since OE and MA do not incorporate shrinkage, their measure responsiveness equals one.

The second measure is risk-standardized mortality rate (RSMR), the current CMS measure for 30-day mortality and 30-day readmissions. CMS now uses three years of claims data for its RSMR calculations, though it initially used one year. The formula for RSMR is:

$$\hat{u}_j^{RSMR} = \frac{\sum_{i=1}^{n_j} f(\hat{\beta}_0 + \hat{\theta}_j + \hat{\beta}_1 X_{ij})}{\sum_{i=1}^{n_j} f(\hat{\beta}_0 + \hat{\beta}_1 X_{ij})} \cdot \bar{y}$$

where $f()$ is the inverse of the logistic link function. For this simulation, $\hat{\beta}_0$, $\hat{\theta}_j$, and $\hat{\beta}_1$ are estimated via a multilevel logistic model with a hospital random effect. The third measure I test is a novel measure that I call the average best linear unbiased estimator (ABLUP). ABLUP, also a shrinkage estimate, is calculated using the same logistic model estimates as RSMR:

$$\hat{u}_j{}^{ABLUP} = \frac{\sum_{i=1}^{N} f(\hat{\beta}_0 + \hat{\theta}_j + \hat{\beta}_1 X_i)}{N}$$

where N is the total number of patients across all hospitals. Thus, ABLUP can be interpreted as the hospital's average of predicted mortality across all possible patients in the sample. Although ABLUP and RSMR are derived from the same logistic model, they do not produce identical estimates, which is apparent when assuming all patients are uniform in their characteristics. In this case, $\hat{u}_j{}^{RSMR} = \hat{u}_j{}^{ABLUP} \frac{\bar{y}}{f(\hat{\beta}_0)}$. The shrinkage factors of RSMR and ABLUP served as estimates of measure responsiveness. These were estimated as $\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\bar{u}(1-\bar{u})}{n_j}}$ and $\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\bar{u}(1-\bar{u})}{n_j}} \cdot \frac{\bar{y}}{f(\hat{\beta}_0)}$, respectively, where $\bar{u}$ is the average of hospitals' observed mortality rates.[6]

The fourth and fifth measures are the Dimick-Staiger measure (DS) (Dimick et al. 2009) and the hierarchical Poisson measure (HP) (Ryan et al. 2012). Unlike the previously described shrinkage estimators, the DS and HP estimators do not shrink all hospitals' observed mortality rates toward a common mortality average. Instead, mortality rates are shrunk toward values that are specific to a hospital's patient volume. Both estimators are calculated according to the following formula:

$$\hat{u}_j{}^{DS,HP} = \bar{u}_j s_j^{DS,HP} + \hat{v}_j^{DS,HP}\left(1 - s_j^{DS,HP}\right)$$

[6] See documentation for the Stata command meqrlogit for details on the calculation of $\hat{\sigma}_u^2$ (StataCorp, 2013).

where $\bar{u}_j$is a hospital's observed mortality, $s_j^{DS,HP}$, is the DS or HP shrinkage factor and $\hat{v}_j^{DS,HP}$ is the hospital's predicted mortality based on its volume. There are several differences between the DS and HP measures regarding how shrinkage weights and volume-predicted mortality are calculated. Unlike for DS, HP estimates of volume-specific mortality are derived from a nonlinear model (a negative binomial model for number of deaths), HP is calculated from hospital-level data rather than patient-level data, and HP uses a maximum likelihood approach to estimate shrinkage weights.[7] As in the case of RSMR and ABLUP, the shrinkage factors employed in DS and HP served as estimates of their responsiveness.

Simulation Results

Figure 3.2 illustrates each 30-day mortality measure's overall performance in terms of accuracy and responsiveness. Note that the horizontal axis is reverse-coded, with greater accuracy measures displayed farther to the right. To consider the magnitude of measurement error in relation to average hospital performance, recall that the average hospital 30-day mortality is 0.209. First, consider the one-year mortality measures, which tend to perform least accurately and with the least responsiveness. OE, the one-year measure without shrinkage, has a substantial amount of error, with a RSME of roughly 0.1. Shrinkage measures perform much more accurately, with RMSE less than 0.06. However, the loss of measure responsiveness entailed by shrinkage estimation is also substantial. The average shrinkage factor facing hospitals ranges from 0.62 to 0.71 for one-year shrinkage measures. This level of

[7] For the details of how volume-predicted mortality and shrinkage factors are calculated for DS and HP, see Dimick et al. (2009), Ryan et al. (2012). For details on adjusting the DS estimator for patient covariates, see Staiger et al. (2009).

measure responsiveness can be viewed as a tax of roughly 30-40 percent on measure improvement. Note that a hospital facing a 0.71 shrinkage factor must improve mortality by 1/.71 = 1.4 percentage points to increase measured mortality by one percentage point.

Table 3.1 contains the main simulation results, which present the measures' overall accuracy and responsiveness by hospital size. Columns (1) and (5), which contain the findings in Figure 3.2, confirm that the shrinkage estimators have greater accuracy and lower measure responsiveness than the estimators without shrinkage, OE and MA. [8] Columns (2) and (5) present RSME and measure responsiveness for hospitals in the bottom quartile of AMI volume. These smaller hospitals experience the greatest improvements in RMSE and greatest reductions in responsiveness when shrinkage estimators are employed. For example, with one year of mortality data, RMSE for the non-shrinkage measure is 0.17, and the shrinkage measure RMSR reduces this error to 0.09. However, RSMR also decreases measure responsiveness from one to 0.36. These differences in the accuracy and responsiveness between shrinkage and non-shrinkage estimates tend to narrow as more years of data are included in measures. However, even with multiple years of data, responsiveness of shrinkage estimates to the performance of small hospitals remains very low, at 0.50 for the three-year RSMR. As shown in columns (4) and (8) of Table 3.1, shrinkage does not appear to reduce error in estimating large hospitals' performance. For larger hospitals, error is slightly greater for measures without shrinkage, and the responsiveness of shrinkage measures ranges from 0.83 to 0.96.

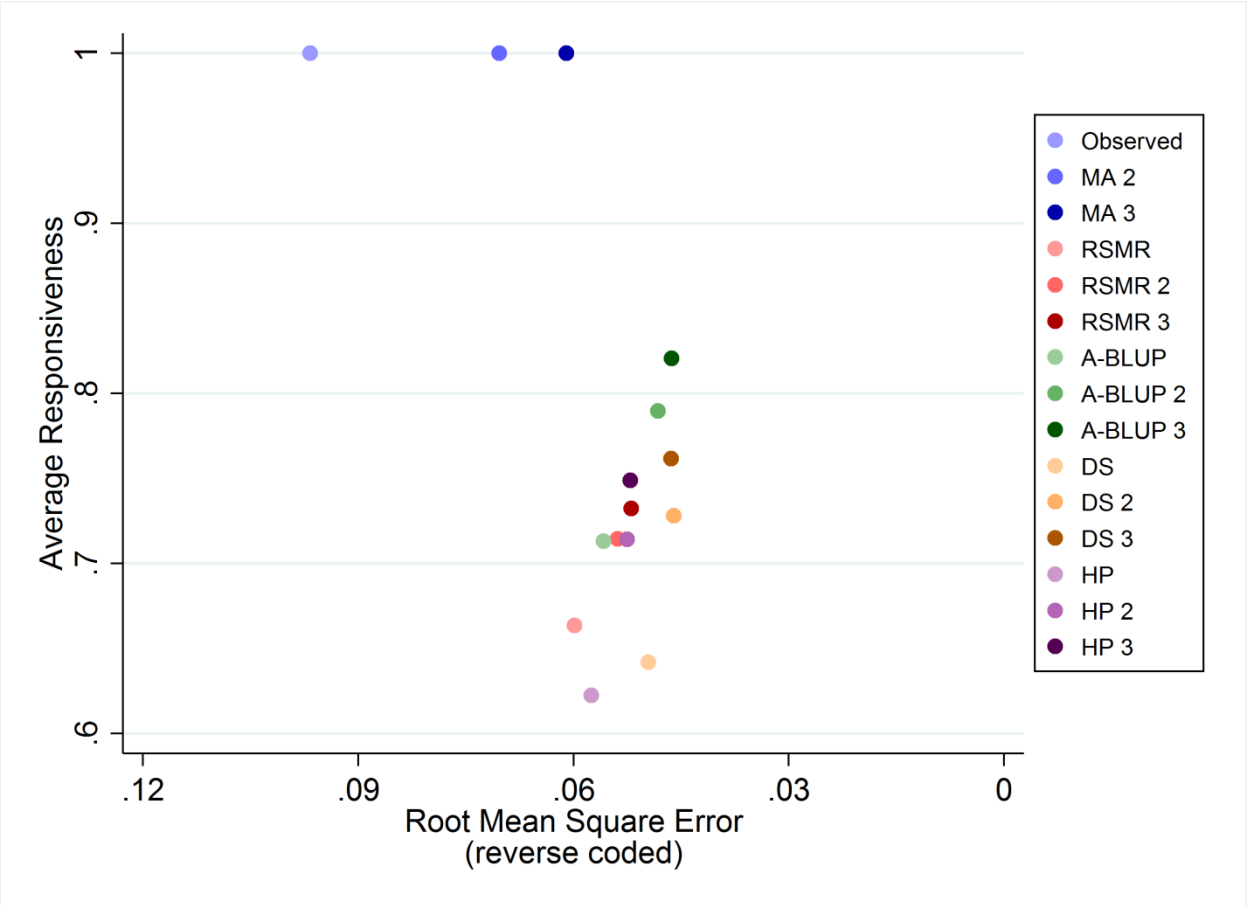
[8] Although several point estimates presented in Table 3.1 are very similar, given the large number of simulation iterations, the differences in measure accuracy and responsiveness between alternate measures tend to be highly statistically significant in paired t-tests.

Figure 3.3 aids in demonstrating the substantial variation in the responsiveness of shrinkage measures by hospital size. The figure presents, from a representative simulation iteration, the responsiveness of one-year shrinkage measures for each decile of hospital AMI volume. Responsiveness increases at a decreasing rate with respect to hospital volume, and there is considerable variation in responsiveness of shrinkage estimators across hospital size. Responsiveness to hospital performance is approximately 0.2 for hospitals below the 10th percentile of AMI volume, and approximately 0.9 for hospitals above the 90th percentile. Since measures only approach full responsiveness asymptotically as sample size increases, measures are not fully responsive to hospital performance for even the largest hospitals in the sample. There is also heterogeneity across shrinkage estimators in terms of their responsiveness. Within each hospital decile, the difference between the most and least responsive measure is 0.11 on average.

Choosing among performance measures does not always entail an accuracy-incentives tradeoff. For all estimators, incorporation of additional years of data tends to improve both measure accuracy and responsiveness. The exception to this pattern is the three-year DS measure, which is less accurate than the two-year DS measure. The non-shrinkage measure experiences an especially pronounced gain in accuracy when the measurement timeframe expands. As column (1) of Table 3.1 shows, RSME for this measure falls from 0.097 to 0.061 when three years of data are used instead of one year. The corresponding change in error for the RSMR shrinkage measure was considerably smaller, from 0.060 to 0.052. Increasing the number of observations also improves the responsiveness of shrinkage estimates. However, even with three-years of data, shrinkage estimates are still approximately 20-25% less responsive than the non-

shrinkage estimates, which are fully responsive regardless of the number of observations.

Figure 3.2 30-Day Mortality Measure Accuracy and Responsiveness



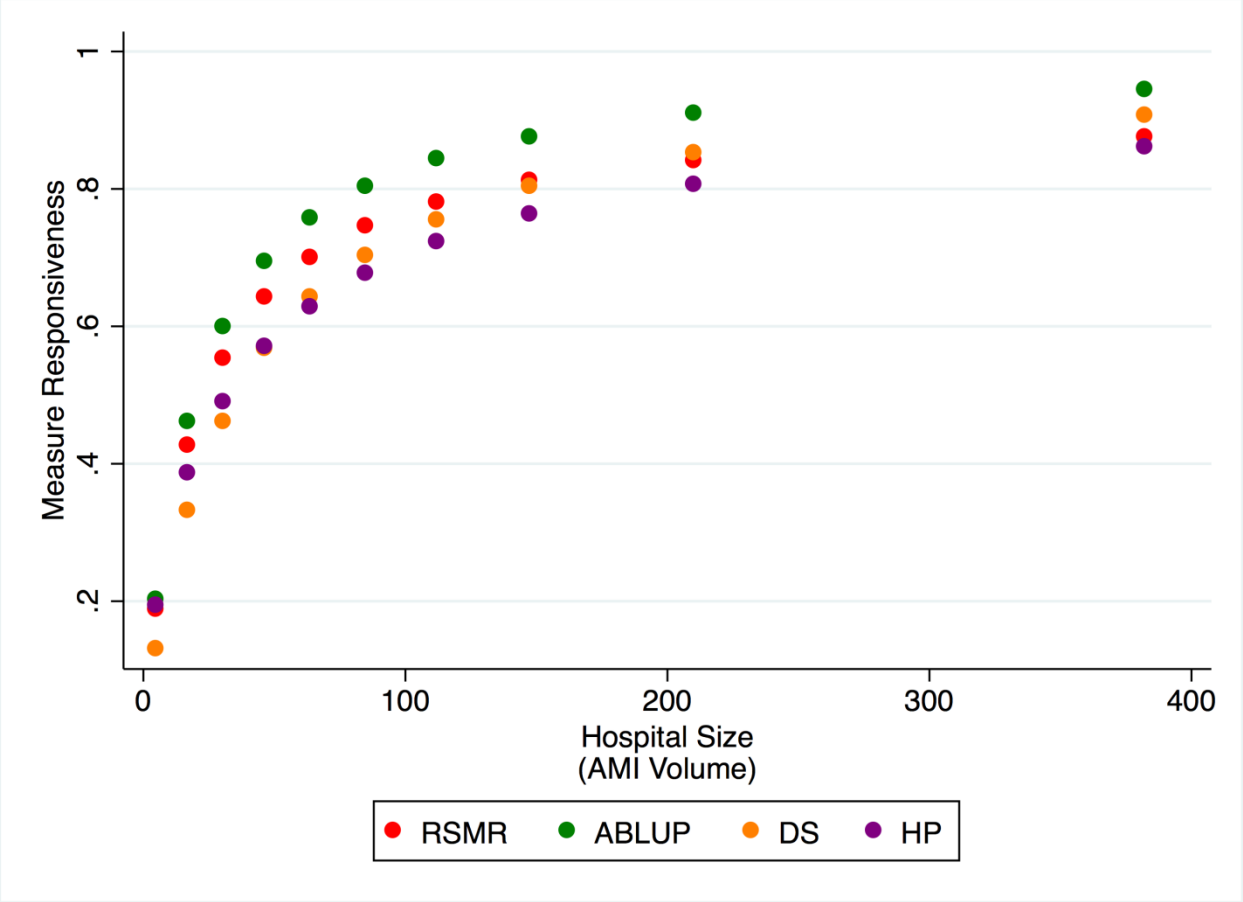
Notes: This figure plots average responsiveness and root mean square error (RMSE) of each hospital 30-day AMI mortality measure, estimated via Monte Carlo simulation with 1000 iterations. Measures incorporate one, two, or three years of prior hospital data (i.e. two years for MA 2). Error is the difference between a measure value and true (unobserved) hospital performance in the following year. A one percentage point difference between a measured and true mortality rate corresponds to MSE of 0.01. Responsiveness is defined as the measure shrinkage factor, which approximates the change in expected measure performance for a change in true performance. Observed over expected (OE) and moving average (MA) are mortality measures without shrinkage. Shrinkage estimators are risk standardized mortality rate (RSMR), average best linear unbiased estimate (ABLUP), Dimick-Staiger (DS) and hierarchical Poisson (HP).

Table 3.1 30-Day Mortality Measure Accuracy and Responsiveness, by Hospital Size

		Root Mean Square Error				Responsiveness			
			By Hospital Size				*By Hospital Size*		
	Estimator	*All Hospitals* (1)	*Small* (2)	*Medium* (3)	*Large* (4)	*All Hospitals* (5)	*Small* (6)	*Medium* (7)	*Large* (8)
One Year	OE*	.0967	.1746	.0513	.0260	1	1	1	1
	RSMR	.0599	.0942	.0473	.0277†	.663	.362	.723	.857
	ABLUP*	.0558	.0863	.0450	.0279	.713	.389	.777	.922
	DS*	.0496	.0708	.0444	.0274	.642	.303	.696	.884
	HP	.0575	.0905	.0447	.0288	.623	.339	.663	.834
Two Years	MA*	.0703	.1245	.0400	.0249	1	1	1	1
	RSMR	.0538	.0843	.0419	.0277†	.714†	.458	.774	.861
	ABLUP*	.0482	.0739	.0385	.0272	.790	.507	.855	.952
	DS**	.0460	.0674	.0393	.0268	.728	.411	.793	.927
	HP	.0525	.0831	.0397	.0280	.714†	.447	.764	.893
Three Years	MA**	.0610	.1038	.0384	.0289	1	1	1	1
	RSMR	.0519	.0791	.0415	.0308†	.732	.502	.789	.859
	ABLUP**	.0463	.0677	.0383	.0308†	.821	.563	.884	.962
	DS	.0464	.0666	.0396	.0305	.762	.462	.827	.941
	HP	.0521	.0804	.0401	.0315	.749	.494	.800	.911

Notes: Cells contain either the root mean squared error (RMSE) or average responsiveness of each hospital 30-day AMI mortality measure, estimated via Monte Carlo simulation with 1000 iterations. Measures incorporate one, two, or three years of prior hospital data. Error is the difference between a measure value and true (unobserved) hospital performance in the following year. Responsiveness is defined as the measure shrinkage factor, which approximates the change in expected measure performance for a change in true performance. A one percentage point difference between a measured and true mortality rate corresponds to MSE of 0.01. Observed over expected (OE) and moving average (MA) are mortality measures without shrinkage. Shrinkage estimators are risk standardized mortality rate (RSMR), average best linear unbiased estimate (ABLUP), Dimick-Staiger (DS) and hierarchical Poisson (HP). Small and large hospitals have annual AMI volume in bottom or top quartile, respectively. Medium hospitals have AMI volume in the middle quartiles. Estimators marked by * are non-dominated on the basis of overall RMSE and responsiveness by other estimators with the same number of years of data. Estimators marked by ** are non-dominated among all estimators regardless of the number of years of data. Within each column, paired t-tests indicate statistically significant ($p<0.05$) differences between all pairwise cell comparisons except for those indicated by †.

Figure 3.3 30-Day Mortality Measure Responsiveness, by Hospital Size



Notes: This figure presents average measure responsiveness within deciles of hospital size. Responsiveness is defined as the measure shrinkage factor. Shrinkage estimators are risk standardized mortality rate (RSMR), average best linear unbiased estimate (ABLUP), Dimick-Staiger (DS) and hierarchical Poisson (HP). Each measure included in this figure uses a single year of mortality data. Data for this figure are drawn from a single representative simulation iteration.

Although additional years of data improved measure accuracy in the simulation, this finding may not generalize to settings in which there is substantial drift in agent behavior over time. If there is extensive drift, early outcomes are less informative of current performance. To demonstrate the sensitivity of measure accuracy to the extent of performance drift, I conduct two secondary simulations. In the first, a no-drift case, each hospital's true mortality rate is fixed over time. In the second, strong-drift case, each hospital has an annual growth rate in mortality improvement (percent change per year) that is drawn from a normal distribution with mean zero and standard deviation of 15%. All other data-generation parameters are the same as in the previously described simulation. In each case, I calculate the RSME of three measures of hospital mortality: one-year observed mortality, three-year mortality average (unweighted), and a three-year weighted average of mortality. Rather than selecting arbitrary weights for the weighted average, I calculate weights for years *t-1*, *t-2,* and *t-3* using constrained linear regression. In each simulation iteration, I regress hospital observed mortality in year *t-1* on observed mortality in years *t-2, t-3, and t-4,* with the constraint that the sum of these coefficients equals one. The resulting coefficients serve as the weights for mortality in years *t-1*, *t-2* and *t-3*, respectively.

Table 3.2 presents the results from these simulations. As shown in column (1), when there is no drift in hospital performance, a moving average has lower RMSE than a one-year estimate. As expected, the constrained regression produced equal weights for all measurement years in this case. As shown in column (2), in the case of substantial performance drift, a three year unweighted average is less accurate than a one-year estimate (0.116 vs 0.109 RMSE). The weighted average, with average weights of 0.67, 0.31 and 0.02 for mortality data from years *t-1*, *t-2*, and *t-3*, outperforms both alternate

measures. Thus, even in the case of changing hospital performance, incorporating early data into measures can increase accuracy if those data are weighted appropriately.

Finally, note that several estimators evaluated in the primary simulation dominate others on the basis of both accuracy and responsiveness. For example, the DS estimator is both more accurate and more responsive than the HP estimator. Similarly, the novel measure ABLUP tends to dominate the current CMS approach, RSMR. The performance frontier of all measures is comprised of the two-year DS, three-year ABLUP, and three-year MA. The RMSE of these measures ranges from 0.061 to 0.046, and the responsiveness ranges from 0.73 to 1. Notably, volume-adjusted shrinkage estimators DS and HP, which shrink observed mortality toward a target that is specific to hospital volume, do not dominate ABLUP and RSMR, which are not volume adjusted. To understand this result, recall that shrinkage measures have greater shrinkage when there is lesser cross-hospital variation in performance. Volume-adjusted shrinkage estimators attribute some hospital performance variation to hospital volume, thereby reducing residual cross-hospital variation increasing shrinkage, and reducing measure responsiveness.

Table 3.2 Comparison of Measure Accuracy for Moving Averages

	Root Mean Square Error	
Estimator	*No Temporal Trend in Performance* (1)	*Temporal Trend in Performance* (2)
One-Year Observed Mortality	.101	.109
Three-Year Unweighted Average	.059	.116
Three-Year Weighted Average	.059	.100
	Moving Average Weights	
Year Before Index Year	*No Temporal Trend in Performance*	*Temporal Trend in Performance*
t-1	0.33†	.67
t-2	0.33†	.31
t-3	0.33†	.02

Notes: This table compares the accuracy of moving averages for performance measurement in two scenarios of hospital performance trajectories. In the column 1 simulation, hospital performance is constant over time. In the column 2 simulation, each hospital improves at a annual rate drawn from a normal distribution with mean zero and standard deviation of 15 percentage points. The Monte Carlo simulations are iterated 1000 times. Error is the difference between a measure value and true (unobserved) hospital performance in the following year, year *t* . The weights for each year of data in 3-year moving averages are determined by constrained linear regression of observed mortality in year *t-1* on observed mortality in years *t-2, t-3* and *t-4,* with coefficients summing to one. According to paired t-tests, within each simulation, all values of RMSE and all weights exhibit statistically significant pairwise differences except for those indicated by †.

3.5 Policy Implications and Conclusion

These theoretic and empirical findings can inform the design of public policies involving performance measurement. The results highlight a substantial tradeoff involved in the choice of performance estimation technique. Although accuracy and responsiveness to agent behavior are both economically desirable features of performance measures, one feature generally comes at the cost of the other. Indeed, in policy settings like health care and education, where ordinary performance estimates are unreliable, shrinkage estimates are least responsive to agent performance. In the case of hospital performance measurement, the magnitude of this loss in responsiveness is economically significant, and may substantially dilute performance incentives. In addition, the magnitude of distortion varies substantially across hospitals, affecting small hospitals to a much greater degree.

As demonstrated in Section 3.2, the appropriate choice of estimation technique depends on a policy's goals. In education, policies that identify inferior teachers for replacement or inferior schools for closure may be welfare-improving even if the policy does not produce a behavioral response. Because the goal of these policies is selecting superior agents rather than incentivizing agent performance, shrinkage estimation seems appropriate. However, for performance payment schemes in which payment is a function of a teacher's absolute performance, shrinkage estimation will tend to dilute incentives unless bonus payments are increased to compensate for reduced measure responsiveness. If performance pay is based on teachers' performance relative to one another, then shrinkage estimation may not distort incentives if teachers face similar

shrinkage factors.[9] However, if there is substantial variation in class sizes, then it will be difficult for teachers in smaller classrooms to receive relative performance bonuses.

Shrinkage estimation may be less appropriate for health care policies, which tend to emphasize incentives. Retaining superior agents (i.e. shutting down inferior hospitals or medical practices) is not a focus of current or proposed initiatives. In the case of performance-based payment to hospitals and physicians, the goal is clearly to incentivize better performance. Shrinkage measurement seems generally inappropriate for performance payment policies like Medicare's hospital readmissions penalties because shrinkage dilutes provider incentives and it is unclear how improved measurement accuracy would contribute to improved welfare. Even if performance payment were based on relative performance, the substantial variation in the size of patient samples across medical providers means that shrinkage estimation could dilute the incentives for providers serving fewer people. The case of public disclosure of quality information is more ambiguous. While publicly disclosing a less responsive performance measure may reduce demand elasticity to provider quality, a more accurate signal could improve patients' choice of hospital. Whether or not to shrink these performance estimates depends on the comparison between the welfare gains from more efficient patient sorting to the welfare gains from increased provider quality spurred by from demand elasticity to quality.

[9] If all teachers have identical numbers of students, then the use of shrinkage estimation does not change the rank order of teacher performance. Moreover, the responsiveness of a teacher's performance *rank* to their true performance is also unchanged by shrinkage estimation. When employing shrinkage estimation, the decrease in the responsiveness of absolute measured performance would be exactly offset by a reduction in the difference between the measured performance of different teachers. Thus, the amount of performance improvement required to increase a teacher's rank would be unchanged.

The simulation results also highlight that some measurement techniques may outperform others with respect to both accuracy and incentives. Policymakers should select measures from this frontier, though it is possible that the relative performance of each technique will vary according to the policy setting. The results also demonstrate that incorporating more observations into performance measures is a substitute to shrinkage estimation in improving measure accuracy. For measures without shrinkage, the gains in accuracy from including more data were considerable. Additional accuracy gains from applying shrinkage may not be worth the loss in measure responsiveness. Even if agent performance changes over time, early data can improve measure accuracy when included in a weighted average of performance.

The analysis in this paper assumes risk-neutrality of agents, which may not hold in all policy settings. A classic finding in the principal-agent literature is that, in determining optimal compensation, agent risk aversion introduces a tradeoff between incentive power and insurance for agents (Gibbons and Roberts, 2013). Although high-powered incentives can still incentivize efficient agent performance, they expose agents to risk. Thus, high-powered incentives may be inappropriate when agents are risk averse. While estimating agent performance with shrinkage does provide some insurance to agents, it is likely to be a blunt tool for this purpose. The shrinkage factors used in performance measurement are not calculated to optimally balance incentives and agent insurance. Thus, even if the optimal incentive power of health care or education policies is not very high (i.e. due to agent risk aversion or multitasking concerns), shrinkage estimation seems unlikely to produce those optimally powered incentives.

Finally, a policymaker's choice of measurement technique may be affected by fairness concerns. An agent may view noisier performance measures as less fair since ratings can vary widely over time even as agent behavior is constant. Similarly, a policymaker may be hesitant to employ a less accurate measurement technique that increases the possibility of type I or type II errors in rewarding or penalizing agents. Alternatively, shrinkage measures may be viewed as less fair. For a given agent, errors from measures without shrinkage will tend to even out over time, while errors from shrinkage estimates are persistent. Shrinkage estimates will persistently underestimate the performance of high-performing agents, and overestimate the performance of low-performing agents. These errors are magnified for agents with fewer observations. Thus, agents may view shrinkage estimation as unfair because their efforts to improve quality performance are not reflected fully in their measured performance.

3.6 References

Arrow, K.J., 1963. Uncertainty and the Welfare Economics of Medical Care. American Economic Review 53, 941–973.

Ash, A., Fienberg, S., Louis, T.A., Normand, S.-L.T., Stukel, T.A., Utts, J., 2012. Statistical Issues in Assessing Hospital Performance.

Burwell, S.M., 2015. Setting value-based payment goals - HHS efforts to Improve U.S. health care. New England Journal of Medicine 372, 897-899.

Chandra, A., Finkelstein, A., Sacarny, A., Syverson, C., 2013. Healthcare Exceptionalism? Productivity and Allocation in the US Healthcare Sector. National Bureau of Economic Research Working Paper Series 19200.

Chetty, R., Friedman, J.N., Rockoff, J.E., 2014. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. American Economic Review 104, 2593–2632.

Dimick, J.B., Staiger, D.O., Baser, O., Birkmeyer, J.D., 2009. Composite measures for predicting surgical mortality in the hospital. Health Affairs 28, 1189–98.

Dimick, J.B., Staiger, D.O., Birkmeyer, J.D., 2010. Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment. Health Services Research 45, 1614–29.

Dimick, J.B., Welch, H.G., Birkmeyer, J.D., 2004. Surgical mortality as an indicator of hospital quality: the problem with small sample size. JAMA 292, 847–51.

Dranove, D., Jin, G.Z., 2010. Quality Disclosure and Certification: Theory and Practice. Journal of Economic Literature 48, 935–963.

Farmer, A., Terrell, D., 1996. Discrimination, Bayesian Updating of Employer Beliefs, and Human Capital Accumulation. Economic Inquiry 34, 204–219.

Gaynor, M., Seider, H., Vogt, W.B., 2005. The Volume–Outcome Effect, Scale Economies, and Learning-by-Doing. American Economic Review: Papers and Proceedings 95, 243–247.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2014. Bayesian Data Analysis, Third Edit. ed. CRC Press.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, p. 258.

Gibbons, R., Roberts, J., 2013. The Handbook of Organizational Economics. Princeton University Press.

Guarino, C., Maxfield, M., Reckase, M.D., Thompson, P., Wooldridge, J.M., 2014. An Evaluation of Empirical Bayes' Estimation of Value-Added Teacher Performance Measures. Michigan State University Education Policy Center Working Paper 31.

Kane, T., Staiger, D., 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. National Bureau of Economic Research Working Paper Series 14607.

Kane, T.J., Staiger, D.O., 2002. The Promise and Pitfalls of Using Imprecise School Accountability Measures. Journal of Economic Perspectives 16, 91–114.

Koedel, C., Mihaly, K., Rockoff, J.E., 2015. Value-added modeling: A review. Economics of Education Review in press.

Kolstad, J., 2013. Information and quality when motivation is intrinsic: evidence from surgeon report cards. American Economic Review 103, 2875–2910.

Krumholz, H.M., Wang, Yun, Mattera, J.A., Wang, Yongfei, Han, L.F., Ingber, M.J., Roman, S., Normand, S.-L.T., 2006. An administrative claims model suitable for profiling

hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation 113, 1683–92.

Morris, C., 1983. Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association 78, 47–55.

Normand, S.-L.T., Shahian, D.M., 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Statistical Science 22, 206–226.

Normand, S.-L.T., Wolf, R.E., Ayanian, J.Z., McNeil, B.J., 2007. Assessing the accuracy of hospital clinical performance measures. Medical Decision Making 27, 9–20.

Nyweide, D.J., Weeks, W.B., Gottlieb, D.J., Casalino, L.P., Fisher, E.S., 2009. Relationship of primary care physicians' patient caseload with measurement of quality and cost performance. JAMA 302, 2444–50.

Phelps, E., 1972. The Statistical Theory of Racism and Sexism. American Economic Review 62, 659-661.

Richardson, S.S., 2013. Integrating pay-for-performance into health care payment systems. Harvard University unpublished dissertation.

Rothstein, J., 2015. Teacher Quality Policy When Supply Matters. American Economic Review 105, 100–130.

Ryan, A., Burgess, J., Strawderman, R., Dimick, J., 2012. What is the best way to estimate hospital quality outcomes? A simulation approach. Health Services Research 27, 1699–718.

Silber, J.H., Rosenbaum, P.R., Brachet, T.J., Ross, R.N., Bressler, L.J., Even-Shoshan, O., Lorch, S. a, Volpp, K.G., 2010. The Hospital Compare mortality model and the volume-outcome relationship. Health Services Research 45, 1148–67.

Skrondal, A., Rabe-Hesketh, S., 2009. Prediction in multilevel generalized linear models. Journal of the Royal Statistical Society Series A 172, 659–687.

Staiger, D., Rockoff, J., 2010. Searching for effective teachers with imperfect information. Journal of Economic Perspectives 24, 97–117.

StataCorp, 2013. Stata 13 Base Reference Manual. Stata Press, College Station, TX.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1954–1955 1, 197–206.

Thomas, J.W., Hofer, T.P., 1999. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. Medical Care 37, 83–92.

Werner, R.M., Bradlow, E.T., 2006. Relationship between Medicare's hospital compare performance measures and mortality rates. Journal of the American Medical Association 296, 2694–702.

Appendices

APPENDIX 1

Measures of Low-Value Services

Services were screened for measure appropriateness from the sources listed in the paper's methods section according to the following criteria: (1) the service must apply to the general Medicare population; (2) appropriate use of the service (if present) must be plausibly distinguishable from wasteful use using procedural and diagnostic codes from the date of service, site of care, beneficiary demographic information, and chronic condition indicators; (3) the evidence base establishing the low-value of the service must have existed prior to 2009. The feasibility denoted by the second criterion was determined by the physicians on our research team. For some services (e.g., imaging for pulmonary embolism without moderate or high pre-test probability), there was an obvious lack of clinical information in claims necessary to define the low-value scenario (e.g., pre-test probability of pulmonary embolism depends in part on heart rate and physical exam findings not recorded in claims). For other services, we inspected a small random sample of claims detected by preliminary measure algorithms to determine if cases of potentially appropriate use could be systematically excluded. The adequacy of information on symptoms in particular determined the inclusion or exclusion of many candidate services. For example, to identify cardiac stress tests for low-risk, asymptomatic patients would require excluding cases with a wide range of symptoms, including non-specific symptoms (e.g., nausea and diaphoresis), as well as cases with risk factors present that may not be captured in claims (e.g., smoking status, family history, dyslipidemia). In contrast, symptoms of carotid artery disease are more circumscribed as they relate directly to transient ischemic attacks and strokes; thus, we

could more confidently exclude appropriate use in developing a measure of screening for asymptomatic carotid artery disease. Similarly, the cardiac stress testing measure we did include in our analysis (for adults with stable coronary disease) depended only on site of service and prior diagnoses, not symptoms, in order to assess appropriateness. We applied the second criterion more leniently when defining more sensitive versions of each measure (e.g. relaxing from USPSTF D recommendation criteria to C criteria).

A primary finding of our study was that the amount of low-value spending detected by our measures varied widely between more sensitive and more specific versions of our measures. For a service to be included in our study, we required that a more specific version – one that convincingly excluded most if not all cases of appropriate use – could be developed. The difference in spending between sensitive and specific versions, however, was not factored into the measure inclusion decision. Indeed, for some services (e.g., vertebroplasty), the sensitive version was quite specific and vice-versa, with little difference in detected utilization between the two. After the final measures had been developed, the six measure categories (low-value cancer screening, etc.) were defined based on service type and measures were assigned to these categories.

In order to ensure that measures detected their target services across clinical settings, measures were developed from both the 2009 Carrier and Outpatient Research Identifiable Files (RIFs). Services provided in hospital outpatient departments or by hospital-employed providers appear either in the Outpatient RIF alone or in both the Outpatient and Carrier RIF, whereas physician and ancillary services provided in inpatient settings or in non-hospital outpatient settings appear in the Carrier RIF. In

both files, claims chronicle services using Current Procedural Terminology codes and document accompanying diagnoses using ICD-9 codes. Additional demographic information necessary for measure development (i.e. age and sex) was obtained from the 2009 enrollment (denominator) file, and summary spending totals and conditions from the Chronic Condition Warehouse (CCW) were obtained from the Beneficiary Annual Summary File. Together, these variables served as the basis for measure development. Because CPT codes are revised annually, appropriate CPT codes were selected based on their definitions as of January 1, 2009. For the development of some measurement algorithms, we also employed the Berenson-Eggers Type of Service (BETOS) coding system for identifying CPT codes in broader clinical categories.

In order to assess for the presence of chronic conditions as of the service date, we employed CCW variables specifying each relevant condition's date of first occurrence. Additional past diagnoses were assessed using ICD-9 codes present in the 2008 and 2009 Carrier and Outpatient RIFs. When measure restriction criteria required assessment of whether certain services preceded or followed a service of interest (e.g. whether a surgical procedure followed a chest x-ray), the relevant preceding or following service was detected using CPT or BETOS codes in the 2008 and 2009 Carrier and Outpatient RIFs. Column two of Table A1.1 lists all relevant CPT, ICD-9 and BETOS codes used for service detection. Emergency department visits were detected according to methods described in a prior study.[1] Inpatient stays were identified based on the presence of claims in the 2008 and 2009 Medicare Provider and Analysis Review (MedPAR) files.

For all measures, standardized prices were calculated as the median of total allowed charges for relevant services. Allowed charges included payments from

Medicare, beneficiaries, and any other payers. For the majority of measures, relevant services were defined to include both the main detected service and specific services frequently delivered as a part of the detected service (e.g. venipuncture with PSA screening). These additional services were included in spending calculations if they occurred on the day of the detected low-value service. We conservatively excluded codes for evaluation and management services (i.e. office visits) from relevant services because they could have occurred even in the absence of the detected service.

Two alternate approaches to defining relevant services were employed for surgical procedures whose complex billing precluded a comprehensive specification of relevant CPT codes. For surgical procedures sometimes occurring in the outpatient setting (renal artery angioplasty or stenting, vertebroplasty/kyphoplasty, and arthroscopic knee surgery), we isolated encounters that appeared in both the Carrier and Outpatient files and totaled all allowed institutional and professional spending that occurred on the day of the detected service across the two files. We examined the most common CPT codes employed on the day of these operations and did not observe any services being delivered that were obviously unrelated to the service of interest. Pricing based on inpatient prospective payments (diagnosis-related groups or DRGs) was avoided when possible because such payments cover a wide array of services that may not be related to the service of interest. However, this approach was necessary for surgical procedures that occurred almost exclusively in the inpatient setting (i.e. carotid endarterectomy and PCI). For these services, prices were determined based on the sum of all spending for services that occurred on the day of the detected services as well as the spending permitted by the DRG for the inpatient stay, obtained from the MedPAR file. In order to limit the inclusion of spending on unrelated services, we restricted the

pricing sample to instances where the detected service was the only procedure listed in the MedPAR stay or where the assigned DRG for the admission corresponded to the detected service. All additional codes used in the pricing of relevant services are listed in column three of Table A1.1

Multiple prices were calculated for measures encompassing multiple services with substantially varied prices. For example, colon cancer screening prices were calculated separately for fecal occult blood testing and other colon cancer screening modalities, and prices for stress testing were calculated separately for exercise treadmill tests with electrocardiographic monitoring and for tests involving advanced imaging modalities.

In order to avoid counting a single service multiple times in frequency or spending calculations, we did not count any detected services that was recorded as having occurred within seven days of the same type of detected service for each beneficiary.

Table A1.1: Codes Used for Measures of Low-Value Services

Measure	Codes for detection and restriction criteria	Additional codes for pricing	Group qualifying
Cancer screening for patients with chronic kidney disease (CKD) receiving dialysis	**BETOS:** P9A P9B (dialysis) **CPT/HCPCS:** 77057 G0202 (breast screening), G0104-G0106 G0120 - G0122 G0328 82270 (colorectal screening), G0102 G0103 84152-84154 (prostate screening), G0101 G0123 G0124 G0141 G0144 G0145 G0147 G0148 P3000 P3001 Q0091 (cervical screening)	**CPT:** 36415 (venepuncture), 77051-77059 (mammography add-on codes), 00810 (endoscopy sedation), 87620-87622 (HPV tests)	Patients with CKD[a]
Cervical cancer screening for women over age 65	**CPT/HCPCS:** G0101 G0123 G0124 G0141 G0144 G0145 G0147 G0148 P3000 P3001 Q0091 (cervical screening) **ICD-9:**180 184x 2190 2331 2332 2333x 6221 (cervical and other relevant cancers, dysplasias) 7950x-7951x (abnormal Papanicolaou finding, human papillomavirus positivity) V1040 V1041 V1322 (history of cervical cancer, other relevant cancers, dysplasia)	**CPT:** 87620-87622 (HPV tests)	Women over 65
Colorectal cancer screening for adults older than age 85 years	**CCW:** Colorectal cancer first indication date **CPT/HCPCS:** 45330-45345 45378-45392 G0104-G0106 G0120-G0122 G0328 82270 (sigmoidoscopy, colonoscopy, barium enema or blood occult test for colon cancer screening)	**CPT:** 00810 (sedation)	Patients over 75
Prostate-specific antigen (PSA) testing for men over age 75	**CCW:** Prostate cancer first indication date **CPT/HCPCS:** G0103 84152-84154 (PSA testing)	**CPT:** 36415 (venepuncture)	Men over 75
Bone mineral density testing at frequent intervals	**CCW:** Osteoporosis first indication date **CPT/HCPCS:** 76977 77078-77080 77083 78350 78351 (bone density testing)	None	Patients with osteoporosis[a]
Homocysteine testing for cardiovascular disease	**CPT/HCPCS:** 83090 (homocysteine chemistry) 82746 82747 82607 (folate or B12 testing) **ICD-9:** 2662 2704 2810-2812 2859 (folate or B12 disorders)	**CPT:** 36415 (venepuncture)	All patients

Table A1.1 (Continued): Codes Used for Measures of Low-Value Services

Hypercoagulability testing for patients with deep vein thrombosis	**CPT/HCPCS:** 83090 85300 85303 85306 85613 86147 (hypercoagulability chemistries) **ICD-9:** 4151 (pulmonary embolism) 4510 45111 45119 4512 45181 4519 4534 (phlebitis, thrombophlebitis and venous embolism of lower extremity vessels) V1251 (history of venous thrombosis and embolism, pulmonary embolism)	**CPT:** 83890-83914 (nucleic acid molecular diagnostics)	Patients with deep vein thrombosis[b]
Parathyroid hormone (PTH) measurement for patients with stage 1-3 CKD	**BETOS:** P9A P9B (dialysis) **CCW:** Chronic kidney disease first indication date **CPT/HCPCS:** 83970 (parathyroid hormone chemistry)	**CPT:** 36415 (venepuncture)	CKD patients[a]
Preoperative chest radiography	**BETOS:** P1x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **CPT/HCPCS:** 71010 71015 71020-71023 71030 71034 71035 (chest x-ray), 19120 19125 47562 47563 49560 58558 (relevant surgical codes not included in BETOS categories)	None	Patients undergoing selected surgeries[b]
Preoperative echocardiography	**BETOS:** P1x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **CPT/HCPCS:** 93303 93304 93306-93308 93312 93315 93318 (echocardiogram) 19120 19125 47562 47563 49560 58558 (relevant surgical codes not included in BETOS categories)	**CPT:** 93303-93352 (echocardiography)	Patients undergoing selected surgeries[b]
Preoperative pulmonary function testing (PFT)	**BETOS:** P1x P2x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **CPT/HCPCS:** 94010 (spirometry)	**CPT:** 94010-94799 (pulmonary non-ventilatory services), 93720-93722 (plethysmography)	Patients undergoing selected surgeries[b]
Preoperative stress testing	**BETOS:** P1x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **CPT/HCPCS:** 78451-78454 78460 78461 78464 78465 78472 78473 78481 78483 78491 78492 93015-93018 93350 93351 (stress testing), 19120 19125 47562 47563 49560 58558 (relevant surgical codes not included in BETOS categories)	**CPT:** 93000-93042 (ECG), 93303-93352 (all echocardiography), 78414-78499 (all cardiovascular nuclear diagnostic), A9500-A9700 (contrast), J0150 J0152 J0280 J1245 J1250 J2785 (pharmacologic stress test injection)	Patients undergoing selected surgeries[b]

Table A1.1 (Continued): Codes Used for Measures of Low-Value Services

Computed tomography (CT) of the sinuses for uncomplicated acute rhinosinusitis	**CPT/HCPCS:** 70486-70488 (CT of maxillofacial area) **ICD-9:** 461x 473x (sinusitis), 2770x 042 07953 279xx (immune disorders), 471x (nasal polyp) 373xx 37600 (eyelid/orbit inflammation), 800xx-804xx 850xx-854xx 870xx-873xx 9590x 910xx 920xx-921xx (head or face trauma)	None	Patients with sinusitis diagnosis[b]
Head imaging in the evaluation of syncope	**CPT/HCPCS:** 70450 70460 70470 70551-70553 (CT or MRI of head or brain) **ICD-9:** 7802 9921 (syncope), 345xx 7803x (epilepsy or convulsions), 43xx (cerebrovascular diseases, including stroke/TIA and subarachnoid hemorrhage), 800xx-804xx 850xx-854xx 870xx-873xx 9590x 910xx 920xx-921xx (head or face trauma), 78097 781xx 7820 7845x (altered mental status, nervous and musculoskeletal system symptoms, including gait abnormality, meningismus, disturbed skin sensation, speech deficits), V1254 V10xx (personal history of stroke/TIA)	None	Patients with syncope diagnosis[b]
Head imaging for uncomplicated headache	**CPT/HCPCS:** 70450 70460 70470 70551-70553 (CT or MRI of head or brain) **ICD-9:** 30781 339xx 364x 7840 (headache or migraine), 33920-33922 33943 (post-traumatic or thunderclap headache), 14xx–208xx 230xx-239xx (cancer), 3463x 3466x (migraine with hemiplegia or infarction), 4465 (giant cell arteritis), 345xx 7803x (epilepsy or convulsions), 43xx (cerebrovascular diseases, including stroke/TIA and subarachnoid hemorrhage), 800xx-804xx 850xx-854xx 870xx-873xx 9590x 910xx 920xx-921xx (head or face trauma), 78097 781xx 7845x (altered mental status, nervous and musculoskeletal system symptoms, including gait abnormality, meningismus, disturbed skin sensation, speech deficits), V1254 V10xx (personal history of stroke/TIA or cancer)	None	Patients with headache diagnosis[b]

Table A1.1 (Continued): Codes Used for Measures of Low-Value Services

Electroencephalogram for headaches	**CPT/HCPCS:** 95812 95813 95816 95819 95822 95827 95830 95957 (electroencephalogram) **ICD-9:** 30781 339xx 346x 7840 (headaches) 345xx 7803x 7810 (epilepsy or convulsions)	None	Patients with headache diagnosis[b]
Back imaging for patients with non-specific low back pain[c]	**CPT/HCPCS:** 72010 72020 72052 72100 72110 72114 72120 72200 72202 72220 72131-72133 72141 72142 72146-72149 72156 72157 72158 (radiologic, CT, and MRI imaging of spine) **ICD-9:** 7213 72190 72210 72252 7226 72293 72402 7242-7246 72470 72471 72479 7385 7393 7394 8460-8463 8468 8469 8472 (back pain, various causes), 14xx–208xx 230xx-239xx (cancer), 800x-839xx 850xx-854xx 86xxx 905xx-909xx 92611 92612 929, 952xx 958xx-959xx (trauma), 3040x-3042x 3044x 3054x-3057x (IV drug abuse), 34460 7292x (neurologic impairment), 4210 4211 4219 (endocarditis), 038xx (septicemia), 01xxx (tuberculosis), 730xx (osteomyelitis), 7806x 7830x 7832x 78079 7808x 2859x (fever, weight loss, malaise, night sweats, anemia not due to blood loss)	None	Patients with back pain[b]
Screening for carotid artery disease in asymptomatic adults	**CPT/HCPCS:** 36222-36224 70498 70547-70549 93880 93882 3100F (carotid imaging) **CCW:** Stroke/TIA first indication date **ICD-9:** 430 431 43301 43311 43321 43331 43381 43391 43400 43401 43410 43411 43490 43491 4350 4351 4353 4358 4359 436 99702 V1254 (stroke/TIA), 3623 36284 (retinal vascular occlusion/ischemia), 7802 781xx 7820 78451 78452 78459 9921 (nervous and musculoskeletal symptoms)	None	All patients

Table A1.1 (Continued): Codes Used for Measures of Low-Value Services

Screening for carotid artery disease for syncope	**CPT/HCPCS:** 36222-36224 70498 70547-70549 93880 93882 3100F (carotid imaging) **CCW:** Stroke/TIA first indication date **ICD-9:** 7802 9921 (syncope), 430 431 43301 43311 43321 43331 43381 43391 43400 43401 43410 43411 43490 43491 4350 4351 4353 4358 4359 436 99702 V1254 (stroke/TIA), 3623 36284 (retinal vascular occlusion/ischemia), 781xx 7820 78451 78452 78459 (nervous and musculoskeletal symptoms)	None	Patients with syncope diagnosis[b]
Stress testing for stable coronary disease	**CPT/HCPCS:** 93015-93018 93350 93351 78451-78454 78460 78461 78464 78465 78472 78473 78481 78483 78491 78492 (stress testing) **CCW:** Ischemic heart disease first indication date, AMI first indication date	**CPT:** 93000-93042 (ECG), 93303-93352 (echocardiography), 78414-78499 (cardiovascular nuclear diagnostic services), A9500-A9700 (contrast), J0150 J0152 J0280 J1245 J1250 J2785 (pharmacologic stress test injection)	IHD patients[a]
Percutaneous coronary intervention with balloon angioplasty or stent placement for stable coronary disease	**CPT/HCPCS:** 92980 92982 (coronary stent placement or balloon angiography) **CCW:** Ischemic heart disease first indication date, AMI first indication date	**DRG:** 246-251[d] (percutaneous cardiovascular procedure)	IHD patients[a]
Renal artery angioplasty or stenting	**CPT/HCPCS:** 35471 35450 37205 37207 75966 75960 (renal artery angioplasty or stenting) **ICD-9:** 4401 40501 40511 40591 (atherosclerosis of renal artery, renovascular hypertension)	None[e]	Patients with hypertension[b]
Carotid endarterectomy in asymptomatic patients	**CPT/HCPCS:** 35301 (carotid endarterectomy) **CCW:** Stroke/TIA first indication date **ICD-9:** 430 431 43301 43311 43321 43331 43381 43391 43400 43401 43410 43411 43490 43491 4350 4351 4353 4358 4359 436 99702 V1254 (stroke/TIA), 3623 36284 (retinal vascular occlusion/ischemia), 781xx 7820 78451 78452 78459 (nervous and musculoskeletal symptoms)	**ICD-9 procedure**: 3812 0040-0042¶ (carotid endarterectomy)	All patients

Table A1.1 (Continued): Codes Used for Measures of Low-Value Services

Inferior vena cava filters for the prevention of pulmonary embolism	**CPT/HCPCS:** 75940 (radiological supervision of inferior vena cava filter placement)	**CPT:** 36010 37620 75825 76937 (catheter insertion, IVC interruption, venography, ultrasound guidance)	All patients
Vertebroplasty or kyphoplasty for osteoporotic vertebral fractures	**CPT/HCPCS:** 22520 22521 22523 22524 (vertebroplasty, kyphoplasty) **ICD-9:** 73313 8052 8054 (vertebral fracture) , 1702 1985 20300-20302 2132 22809 2380 2386 2392 (primary or secondary neoplasm of vertebral column, multiple myeloma, hemangioma)	None[e]	Patients with osteoporosis[a]
Arthroscopic surgery for knee osteoarthritis	**CPT/HCPCS:** 29877 29879 G0289 (knee arthroscopy with chondroplasty) **ICD-9:** 7177 73392 71500 71509 71510 71516 71526 71536 71596 (chondromalacia, osteoarthritis), 8360-8362 7170 71741 (meniscal tear)	None[e]	Patients with arthritis[a]

[a] Defined by presence of CCW first indication date prior to January 1, 2010

[b] Defined by presence of relevant diagnosis or procedure codes during 2009.

[c] We follow prior literature in defining this measure.[2]

[d] The pricing sample was restricted to detected hospital admissions with these DRG codes. All professional charges for expenses incurred on the same day of service were included in pricing estimates.

[e] The pricing sample was restricted to detected episodes that appeared in both the Carrier and Outpatient files. All professional charges for expenses incurred on the same day of service were included in pricing estimates.

[f] The pricing sample was restricted to detected hospital admissions with no procedures besides those listed here. All professional charges for expenses incurred on the same day of service were included in pricing estimates

Primary Analysis

Utilization rates and associated spending for services detected by low-value care measures, presented graphically in Figure 1.1, are presented in tabular form in Table A1.2. Several variables included in our regression analyses merit additional explanation. In order to account for case mix, we included an extensive set of patient characteristics in regressions. These included indicators for 21 CCW diagnoses present before 2009 (derived from claims dating back to 1999) and indicators of having multiple comorbid conditions (2 to 7+). In addition to these variables, we developed indicators for demographic characteristics and clinical conditions qualifying beneficiaries for potential receipt of low-value services, listed in column 4 of Table A1.1. Although these indicators "qualify" beneficiaries for the receipt of services, the indicators do not imply that the receipt of services is appropriate. Instead, the indicators highlight those patients whose characteristics make them eligible to receive a low-value service. For instance, because our measure of low-value PSA testing applies to men over age 75, men over age 75 are the qualifying group for this measure. In our analyses, inclusion of these indicators helps prevent apparent correlations from arising that are driven by the geographic distribution of patients who qualify for low-value services. For instance, if some regions had a higher incidence of both syncope and osteoarthritis of the knee than average and therefore higher population rates of imaging for syncope and arthroscopy knee surgery, without adjustment for the prevalence of syncope and osteoarthritis of the knee, the estimated correlation between these two types of services could be positive even if practice patterns in these regions were the same (or even more conservative) relative to other regions. Notably, our results were not sensitive to the inclusion of these indicators.

Supplementary Analysis

In order to assess whether greater total spending predicts greater measured overuse, we examined the association between regional spending on low-value services and total regional spending for Medicare beneficiaries as a supplementary analysis. To do so, we fitted a linear regression model predicting spending on low-value services for each beneficiary as a function of 2009 mean price-adjusted Medicare Part A and B spending per beneficiary at the HRR level and the same set of beneficiaries' sociodemographic and clinical characteristics included in our primary analysis. To facilitate interpretation, we specified total regional spending per beneficiary in quartiles. Following regression analysis, the statistical significance of the association between spending on low-value services and quartile of overall spending was assessed via Wald test of the null hypothesis that adjusted spending on low-value services was equal across quartiles. Regional total Medicare spending was positively associated with measured low-value spending ($P<0.001$ for test of equality across quartiles). Adjusted per beneficiary spending on services detected by low-value measures ranged from $282 in the lowest quartile of overall spending to $326 in the highest quartile of overall spending. This finding is consistent with the interpretation that variation in total spending is predictive of wasteful practices. However, low-value spending varied by less than 20% across quartiles of total regional spending.

We conducted a sensitivity analysis assessing the association between spending on low-value services and an alternate measure of total regional Medicare spending. The purpose of this analysis was to test whether the inclusion of low-value spending in measures of overall spending induced the positive association presented above. Unlike the analysis, which used a price-adjusted regional measure of overall Part A and Part B

Medicare spending obtained from the Dartmouth Atlas of Health Care, this sensitivity analysis used a measure of overall spending that excluded spending on measured low-value services. The alternate measure was constructed by calculating total Part A and B payments for each beneficiary in our study from the 2009 Beneficiary Annual Summary File (payments by Medicare, beneficiaries, and other payers), multiplying the totals by Dartmouth Atlas regional price adjusters (each calculated as the ratio of price-adjusted regional spending estimates over unadjusted regional spending estimates), subtracting each individual's spending on measured low-value services (based on standardized prices), and computing the average of the resulting value by HRR. The alternate measure of regional total Medicare spending was also positively associated with measured low-value spending ($P<0.001$ for test of equality across quartiles) and the association was not appreciably attenuated by use of the alternate measure. Adjusted per beneficiary spending on services detected by low-value measures ranged from $282 in the lowest quartile of overall spending to $322 in the highest quartile of overall spending.

Table A1.2 Use and Associated Spending of Services Detected by Low-Value Service Measures, by Category

	More Sensitive Measure						More Specific Measure					
Measure Category	Count per 100 bene[a]	% of low-value count	% of benes affected	Spending ($M)	% of low-value spending	% of overall spending[b]	Count per 100 bene[a]	% of low-value count	% of benes affected	Spending ($M)	% of low-value spending	% of overall spending[b]
Cancer Screening	27.0	34%	20%	794	9%	0.26%	10.3	31%	10%	142	7%	0.05%
Diagnostic and preventive testing	11.0	14%	5%	174	2%	0.06%	4.8	14%	3%	77	4%	0.02%
Preoperative testing	7.1	9%	6%	315	4%	0.10%	2.3	7%	2%	125	6%	0.04%
Imaging	25.5	32%	18%	939	11%	0.30%	14.5	43%	12%	620	32%	0.20%
Cardiovascular testing and procedures	9.3	12%	8%	5,886	70%	1.90%	1.2	4%	1%	717	37%	0.23%
Other surgery	0.5	1%	0%	343	4%	0.11%	0.4	1%	0%	259	13%	0.08%
Total	80	100%	42%[c]	8,451	100%	2.73%	33	100%	25%[c]	1,941	100%	0.63%

[a] Count refers to the number of unique incidences of service provision.

[b] Overall spending refers to annual spending for services covered by Part A and B of Medicare. See Table 1.1 for service category assignments and for operational definitions of all measures.

[c] Total does not equal column sum because some patients received multiple different services.

Additional References

1. Colla CH, Wennberg DE, Meara E, et al. Spending differences associated with the Medicare Physician Group Practice Demonstration. *JAMA* 2012: 308(10):1015–23
2. Pham HH, Landon BE, Reschovsky JD, Wu B, Schrag D. Rapidity and modality of imaging for acute low back pain in elderly patients. *Arch Intern Med* 2009;169(10):972–81

APPENDIX 2

Service Detection

This section briefly describes our method for detecting services meeting our operational definitions of low-value service. These methods are described fully in prior work.[1] To detect each service, we first searched for potential low-value services using their *Current Procedural Terminology* (CPT) code in the Medicare Carrier and Outpatient Research Identifiable Files. Services performed in the inpatient setting appear in the Carrier file, which contains claims filed on behalf of physicians and other non-institutional providers. Services performed in the outpatient setting appear in the Carrier and/or the Outpatient file depending on whether they took place in a hospital or non-hospital outpatient setting.

We determined whether target services satisfied our operational definitions of low-value services on the basis of patient demographic and clinical data found in claims or other Medicare research files. For example, the Beneficiary Annual Summary File was the source of patient data on age, sex, and the presence of the chronic conditions available in the Chronic Conditions Warehouse (CCW) segment of the file. In assessing whether a service met the operational definition of a low-value service, we employed claims data from as early as January 1 of the year prior to the service being evaluated. For example, we searched for relevant patient diagnoses on the basis of International Classification of Diseases, Ninth Revision (ICD-9) codes in the claims for target services and in prior claims.

Some operational definitions of low-value services included criteria based on the site of care or the timing of the service. For example, low-value preoperative services were defined as occurring prior to surgical operations, which were detected on the

basis of a CPT code or Berenson Eggers Type of Service (BETOS) code. For some measures, we assessed whether the service occurred during or close to an inpatient stay using the admissions and discharge dates in the Medicare Provider and Analysis Review (MedPAR) files. Similarly, for some measures, we assessed whether the service occurred close to emergency department visit on the basis of emergency department evaluation and management CPT codes in the Carrier and Outpatient files, emergency department revenue center codes in the Outpatient file, and any indication of an emergency department visit in a MedPAR records (i.e. emergency admissions type, emergency room admissions source, or emergency department charges).

In order to avoid detecting the same service twice, we excluded the detection of any low-value service that occurred within seven days of the same type of low-value service. All codes used to detect services are presented in Table A2.1

Table A2.1: Codes for Measures of Low-Value Care

Measure	Codes for detection and restriction criteria	Added pricing codes
Cancer screening for patients with chronic kidney disease (CKD) receiving dialysis	**BETOS:** P9A P9B (dialysis) **CPT:** 77057 G0202 (breast screening), G0104-G0106 G0120 -G0122 G0328 82270 (colorectal screening), G0102 G0103 84152-84154 (prostate screening), G0101 G0123 G0124 G0141 G0143 G0144 G0145 G0147 G0148 P3000 P3001 Q0091 (cervical screening)	**CPT**: 36415 (venepuncture), 77051-77059 (mammography add-on codes), 00810 (endoscopy sedation), 87620-87622 (HPV tests)
Cervical cancer screening for women over age 65	**CPT:** G0123 G0124 G0141 G0143 G0144 G0145 G0147 G0148 P3000 P3001 Q0091 (cervical screening) **ICD-9:**180 184x 2190 2331 2332 2333x 6221 (cervical and other relevant cancers, dysplasias) 7950x-7951x (abnormal Papanicolaou finding, human papillomavirus positivity) V1040 V1041 V1322 V1589 (history of cervical cancer, other relevant cancers, dysplasia)	**CPT:** 87620-87622 (HPV tests)
Colorectal cancer screening for adults older than age 85 years	**CCW:** Colorectal cancer first indication date **ICD-9:** V7651 (colon cancer screening) **CPT:** G0104-G0106 G0120-G0122 G0328 82270 (screening codes for sigmoidoscopy, colonoscopy, barium enema or blood occult test)	**CPT:** 00810 (sedation)
Prostate-specific antigen (PSA) testing for men over age 75	**CCW:** Prostate cancer first indication date **CPT:** G0103 84152-84154 (PSA testing)	**CPT:** 36415 (venepuncture)
Bone mineral density testing at frequent intervals	**CCW:** Osteoporosis first indication date **CPT:** 76070 76071 76075 76076 76078 76977 77078-77081 77083 78350 78351 (bone density testing)	None
Homocysteine testing for cardiovascular disease	**CPT:** 83090 (homocysteine chemistry) 82746 82747 82607 (folate or B12 testing) **ICD-9:** 2662 2704 2810-2812 2859 (folate or B12 disorders)	**CPT:** 36415 (venepuncture)
Hypercoagulability testing for patients with deep vein thrombosis	**CPT:** 81240 81241 83090 85300 85303 85306 85613 86147 (hypercoagulability chemistries) **ICD-9:** 4151 (pulmonary embolism) 4510 45111 45119 4512 45181 4519 4534 4535 (phlebitis, thrombophlebitis and venous embolism of lower extremity vessels) V1251 V1255 (history of venous thrombosis and embolism, pulmonary embolism)	**CPT:** 83890-83914 (nucleic acid molecular diagnostics)

Table A2.1 (Continued): Codes for Measures of Low-Value Care

Parathyroid hormone (PTH) measurement for patients with stage 1-3 CKD	**BETOS:** P9A P9B (dialysis) **CCW:** Chronic kidney disease first indication date **CPT:** 83970 (parathyroid hormone chemistry)	**CPT:** 36415 (venepuncture)
Total or free T3 level testing for patients with hypothyroidism	**CPT:** 84480 84481 (total or free T3) **CCW**: Hypothyroidism first indicaiton date	None
1,25-dihydroxyvitamin D testing in the absence of hypercalcemia or decreased kidney function	**CPT:**82652 (1, 25 dihydroxyvitamin D3) **CCW:** Chronic kidney disease first indication date **ICD-9:** 27542 (hypercalcemia) 58881 (secondary hyperparathyroidism of renal origin) 135x 01x 173x 174x 175x 1890 1891 188x 1830 200x-208x (sarcoidosis, TB, select neoplasms)	None
Preoperative chest radiography	**BETOS:** P1x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **ICD-9** V7281 V7282 V7283 V7284 (preoperative examination **CPT:** 71010 71015 71020-71023 71030 71034 71035 (chest x-ray), 19120 19125 47562 47563 49560 58558 (relevant surgical codes not included in BETOS categories)	None
Preoperative echocardiography	**BETOS:** P1x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **CPT:** 93303 93304 93306-93308 93312 93315 93318 (echocardiogram) 19120 19125 47562 47563 49560 58558 (relevant surgical codes not included in BETOS categories)	**CPT:** 93303-93352 (echocardiography)
Preoperative pulmonary function testing (PFT)	**BETOS:** P1x P2x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **CPT:** 94010 (spirometry)	**CPT:** 94010-94799 (pulmonary non-ventilatory services), 93720-93722 (plethysmography)

Table A2.1 (Continued): Codes for Measures of Low-Value Care

Preoperative stress testing	**BETOS:** P1x P3D P4A P4B P4C P5C P5D P8A P8G (selected surgeries) **CPT:** 75552-75564 75574 78451-78454 78460 78461 78464 78465 78472 78473 78481 78483 78491 78492 93015-93018 93350 93351 0146T 0147T 0148T 0149T (stress testing, cardiac MRI, CT angiography) 19120 19125 47562 47563 49560 58558 (relevant surgical codes not included in BETOS categories)	**CPT:** 93000-93042 (ECG), 93303-93352 (echocardiography), 78414-78499 (cardiovascular nuclear diagnostic services), 75552-75564 (cardiac MRI), 75571-75574 (cardiac CT), A9500-A9700 (contrast), J0150 J0152 J0280 J1245 J1250 J2785 (pharmacologic stress test injection)
Computed tomography (CT) of the sinuses for uncomplicated acute rhinosinusitis	**CPT:** 70486-70488 (CT of maxillofacial area) **ICD-9:** 461x 473x (sinusitis), 2770x 042 07953 279xx (immune disorders), 471x (nasal polyp) 373xx 37600 (eyelid/orbit inflammation), 800xx-804xx 850xx-854xx 870xx-873xx 9590x 910xx 920xx-921xx (head or face trauma)	None
Head imaging in the evaluation of syncope	**CPT:** 70450 70460 70470 70551-70553 (CT or MRI of head or brain) **ICD-9:** 7802 9921 (syncope), 345xx 7803x (epilepsy or convulsions), 43xx (cerebrovascular diseases, including stroke/TIA and subarachnoid hemorrhage), 800xx-804xx 850xx-854xx 870xx-873xx 9590x 910xx 920xx-921xx (head or face trauma), 78097 781xx 7820 7845x (altered mental status, nervous and musculoskeletal system symptoms, including gait abnormality, meningismus, disturbed skin sensation, speech deficits), V1254 V10xx (personal history of stroke/TIA)	None
Head imaging for uncomplicated headache	**CPT:** 70450 70460 70470 70551-70553 (CT or MRI of head or brain) **ICD-9:** 30781 339xx 346x 7840 (headache or migraine), 33920-33922 33943 (post-traumatic or thunderclap headache), 14xx–208xx 230xx-239xx (cancer), 3463x 3466x (migraine with hemiplegia or infarction), 4465 (giant cell arteritis), 345xx 7803x (epilepsy or convulsions), 43xx (cerebrovascular diseases, including stroke/TIA and subarachnoid hemorrhage), 800xx-804xx 850xx-854xx 870xx-873xx 9590x 910xx 920xx-921xx (head or face trauma), 78097 781xx 7820 7845x 79953 (altered mental status, nervous and musculoskeletal system symptoms, including gait abnormality, meningismus, disturbed skin sensation, speech deficits), V1254 V10xx (personal history of stroke/TIA or cancer)	None

Table A2.1 (Continued): Codes for Measures of Low-Value Care

Electroencephalogram for headaches	**CPT:** 95812 95813 95816 95819 95822 95827 95830 95957 (electroencephalogram) **ICD-9:** 30781 339x 346x 7840 (headaches) 345xx 7803x 7810 (epilepsy or convulsions)	None
Back imaging for patients with non-specific low back pain	**CPT:** 72010 72020 72052 72100 72110 72114 72120 72200 72202 72220 72131-72133 72141 72142 72146-72149 72156 72157 72158 (radiologic, CT, and MRI imaging of spine) **ICD-9:** 7213 72190 72210 72252 7226 72293 72402 7242-7246 72470 72471 72479 7385 7393 7394 846x 8472 (back pain, various causes), 14xx–208xx 230xx-239xx (cancer), 800x-839xx 850xx-854xx 86xxx 905xx-909xx 92611 92612 929, 952xx 958xx-959xx (trauma), 3040x-3042x 3044x 3054x-3057x (IV drug abuse), 34460 7292x (neurologic impairment), 4210 4211 4219 (endocarditis), 038xx (septicemia), 01xxx (tuberculosis), 730xx (osteomyelitis), 7806x 7830x 7832x 78079 7808x 2859x (fever, weight loss, malaise, night sweats, anemia not due to blood loss) 72142 72191 72270 72273 7244 (myelopathy, neuritis and radiculopathy)	None
Screening for carotid artery disease in asymptomatic adults	**CPT:** 70498 70547-70549 93880 93882 3100F (carotid imaging) **CCW:** Stroke/TIA first indication date **ICD-9:** 430 431 43301 43311 43321 43331 43381 43391 43400 43401 43410 43411 43490 43491 4350 4351 4353 4358 4359 436 99702 V1254 (stroke/TIA), 3623 36284 (retinal vascular occlusion/ischemia), 7802 781xx 7820 78451 78452 78459 9921 (nervous and musculoskeletal symptoms)	None
Screening for carotid artery disease for syncope	**CPT:** 70498 70547-70549 93880 93882 3100F (carotid imaging) **CCW:** Stroke/TIA first indication date **ICD-9:** 7802 9921 (syncope), 430 431 43301 43311 43321 43331 43381 43391 43400 43401 43410 43411 43490 43491 4350 4351 4353 4358 4359 436 99702 V1254 (stroke/TIA), 3623 36284 (retinal vascular occlusion/ischemia), 781xx 7820 7845x 78459 (nervous and musculoskeletal symptoms)	None
Imaging for diagnosis of plantar fasciitis/heel pain	**CPT:**73620 73630 73650 (foot radiograph) 73718 73719 73720 (foot MRI) 76880 76881 76882 (extremity ultrasound) **ICD-9:**72871 7294 (plantar fasciitis), 71947 7295 (foot pain)	None

Table A2.1 (Continued): Codes for Measures of Low-Value Care

Stress testing for stable coronary disease	**CPT:** 75552-75564 75574 78451-78454 78460 78461 78464 78465 78472 78473 78481 78483 78491 78492 93015-93018 93350 93351 0146T 0147T 0148T 0149T (stress testing, cardiac MRI, CT angiography) **CCW:** Ischemic heart disease first indication date, AMI first indication date	**CPT:** 93000-93042 (ECG), 93303-93352 (echocardiography), 78414-78499 (cardiovascular nuclear diagnostic services), 75552-75564 (cardiac MRI), 75571-75574 (cardiac CT), A9500-A9700 (contrast), J0150 J0152 J0280 J1245 J1250 J2785 (pharmacologic stress test injection)
Percutaneous coronary intervention with balloon angioplasty or stent placement for stable coronary disease	**CPT:** 92980 92982 (coronary stent placement or balloon angiography) **CCW:** Ischemic heart disease first indication date, AMI first indication date	**DRG:** 246-251[a] (percutaneous cardiovascular procedure)
Renal artery angioplasty or stenting	**CPT:** 35471 35450 37205 37207 75960 75966 (renal artery angioplasty or stenting) **ICD-9:** 4401 40501 40511 40591 (atherosclerosis of renal artery, renovascular hypertension), 36221 40xxx 4372 (hypertension)	None[b]
Carotid endarterectomy in asymptomatic patients	**CPT:** 35301 (carotid endarterectomy) **CCW:** Stroke/TIA first indication date **ICD-9:** 430 431 43301 43311 43321 43331 43381 43391 43400 43401 43410 43411 43490 43491 4350 4351 4353 4358 4359 436 99702 V1254 (stroke/TIA), 3623 36284 (retinal vascular occlusion/ischemia), 781xx 7820 7845x (nervous and musculoskeletal symptoms)	**ICD-9 Procedure**: 3812 0040-0042[a] (carotid endarterectomy)
Inferior vena cava filters for the prevention of pulmonary embolism	**CPT:** 37191 37192 (IVC placement, repositioning) 75940 (radiological supervision of inferior vena cava filter placement)	**CPT:** 36010 37620 75825 76937 (catheter insertion, IVC interruption, venography, ultrasound guidance)

Table A2.1 (Continued): Codes for Measures of Low-Value Care

Pulmonary Artery Catheterization in the ICU	**CPT**: 93503 (Swan-Ganz placement) **ICD-9:** 4233 (cardiac tamponade) 4160 4161 4162 4168 4169 (pulmonary hypertension) **MS-DRGs (2008-2012)[d]:** 001-003 005-008 010 020-033 037-042 113-117 129-139 163-168 215-245 252-264 326-358 405-425 453-517 820-830 853-858 876 901-909 927-929 939-941 955-959 969-970 981-989	None
Vertebroplasty or kyphoplasty for osteoporotic vertebral fractures	**CPT:** 22520 22521 22523 22524 (vertebroplasty, kyphoplasty) **ICD-9:** 73313 8052 8054 (vertebral fracture) , 1702 1985 20973 20300-20302 2132 22809 2380 2386 2392 (primary or secondary neoplasm of vertebral column, multiple myeloma, hemangioma)	None[b]
Arthroscopic surgery for knee osteoarthritis	**CPT:** 29877 29879 29880 29881 G0289 (knee arthroscopy with chondroplasty) **ICD-9:** 7177 73392 71500 71509 71510 71516 71526 71536 71596 (chondromalacia, osteoarthritis), 8360-8362 7170 71741 (meniscal tear)	None[b]
Spinal injection for low-back pain	**CPT:** 62311 64483 (epidural injections) 20552 20553 (trigger point injections) 64493 64475 (facet injections) J1438 (etanercept injection) **ICD-9:** 72142 72210 72270 72273 7243 7244 (back pain with radiculopathy) 7213 72190 72210 7222 72252 7226 72280 72283 72293 72400 72402 72403 7242 7245 7246 72470 72471 72479 7384 7385 7393 7384 7385 7393 7394 75612 8460-8463 8468 8469 8472 (other back pain)	None[b]

[a.] The pricing sample was restricted to detected hospital admissions with these DRG codes. All professional charges for expenses incurred on the same day of service were included in pricing estimates along with the DRG allowed charges.

[b] The pricing sample was restricted to detected episodes that appeared in both the Carrier and Outpatient files. All institutional and professional charges for expenses incurred on the same day of service were included in pricing estimates.

[c] The pricing sample was restricted to detected hospital admissions with no procedures besides those listed here. All professional charges for expenses incurred on the same day of service were included in pricing estimates.

[d] Non-medical DRGs were defined according to methods presented in a prior study.[6]

Statistical Analysis: Variation, Consistency and Persistence

In this section, we provide a more detailed description of our methods for producing estimates of variation, consistency, and persistence of organizations' use of low-value services. All analyses involved three general steps. First, we adjusted beneficiary's use of each of the 31 low-value services for case mix. For each service, this case mix adjustment allowed us to isolate each beneficiary's residual low-value service use that could not be explained by various sociodemographic, clinical, temporal and regional characteristics. Second, residuals from the case-mix adjustment model were used to calculate organizations' composite scores of low-value service use. The composite scores reflected the use of several different low-value services, either across the whole study period (i.e. for our analyses of variation and consistency) or during a single year (i.e. for our analysis of persistence). Third, we fit random effects models to these composite scores in order to estimate the parameters of interest. We begin by describing our analysis of organizations' variation in low-value service use. Each subsequent section describes the extensions to these methods that were required for analyzing consistency and persistence of organizational behavior.

Variation

This analysis produced an estimate of the across-organization standard deviation in the case mix-adjusted count of low-value services per 100 beneficiaries. We also present a corresponding estimate of the ratio of adjusted low-value service counts at organizations at the 90th vs 10th percentile.

We performed case mix adjustment via ordinary least squares regressions for each low-value service. In these models, the outcome variable was each the number of times a beneficiary received the low-value service during the year. The regressions included patient sociodemographic characteristics, indicators for patient HRR, and indicators for year. Only beneficiaries who satisfied the denominator criteria for the service were included in the model, since other beneficiaries could not have received the measured service. Because some of the 306 HRRs might be served by only one of the 250 ACOs, we included an additional group of beneficiaries in these models to serve as a regional control group. These additional beneficiaries (n=20,520,493) were not assigned to ACOs, and accumulated the majority of their annual allowed charges for primary care at a non-ACO TIN. Including them in the case mix adjustment models ensured that patient region and provider organization were not perfectly correlated. Regressions were of the following form:

$$Y_{ijkt} = \beta_0 + \beta_1 Covariates_{it} + \beta_2 Year\_indicators_t + \epsilon_{ijkt}$$

In this equation, *i* denotes beneficiary, *j* denotes their assigned provider organization, *t* denotes year, and *k* denotes the service. $Covariates_{it}$ includes the patient sociodemographic and clinical characteristics listed in the manuscript, as well as HRR indicators. We performed these regressions both with and without HRR indicators in order to compare estimates of organizational variation that included adjustment for region to estimates that did not.

The prediction errors from the case mix adjustment models, ϵ_{ijtk}, served as the basis for calculating organizations' composite score for the total number low-

value services per beneficiary. For each organization in our sample, a component measure of each low-value service, $\hat{r}_{jk}$, was calculated as the average residual for beneficiary-years attributed to that organization:

$$\hat{r}_{jk} = \sum_{it} \epsilon_{ijkt} \,/\, \sum_{i} x_{ijk}$$

where x_{ijk} is the number of years during which a beneficiary was assigned to organization *j* and satisfied the denominator criteria for service *k*. $\hat{r}_{jk}$ represents the difference between an organization's average number of low-value services per denominator beneficiary and the number that would be predicted from the case mix adjustment model.

The composite measure for overall low-value service use was calculated as a weighted sum of these component scores:

$$\widehat{R}_J = \sum_k w_k \, \hat{r}_{jk}$$

where w_k is the proportion of all person-year observations in the sample that satisfied the denominator conditions. This approach does not give greater weight in the composite measure to services that apply to a greater proportion of the population. Instead, the weighting standardizes each service's contribution by the number of beneficiaries included in the case mix adjustment models. For intuition behind this result, note that the total number of low-value services per beneficiary would be the same if a service were used one time per person in an entire population or if a service were used twice per person in half of that population. Thus, $\widehat{R}_J$ approximates an organization's residual case mix-adjusted count of all low-value services.

We estimate the across-organization variation in R_j using Fay–Herriot-type models.[2–4] This class of models allows for the estimation of multilevel model parameters after collapsing data to the highest level of analysis, which in our case, is the organization. This approach had computational advantages for our study given that the data contain millions or tens of millions of beneficiary-year observations among hundreds or thousands of organizations. Specifically, for our analysis of organizational variation, we fit the following model:

$$\widehat{R_j} = R_j + e_j$$

$$\text{with } R_j \sim N\left(\mu, \sigma^{(R)}\right) \text{ and } e_j \sim N\left(0, \sigma_j^{(e)}\right)$$

Our parameter of interest is $\sigma^{(R)}$, the across-organization standard deviation of R_j. The purpose of random effects estimation in this context is to account for sampling error, which results in over-dispersion of observed $\widehat{R_j}$ relative to its true distribution. Because, in our Fay–Herriot model, the data are aggregated to the organization level, with a single observation per organization, accounting for sampling error requires separately estimating the sampling variance.

Following methods described for analyzing composite quality measures from the Consumer Assessment of Healthcare Providers and Systems (CAHPS),[5] we calculated the sampling variance of $\widehat{R_j}$ in two steps. First, we calculated the following error term:

$$\varepsilon_{ijk} = \frac{\sum_t \epsilon_{ijkt} - x_{ijk}\,\hat{r}_{jk}}{\sum_i x_{ijk}}$$

When an individual did not qualify for the denominator of a service, ϵ_{ijkt} and x_{ijk} were both zero. The denominator is an organization's total number of

beneficiary-year observations for service *k*. By collapsing the prediction error to the beneficiary-organization-service level, our variance estimate accounts for possible temporal autocorrelation in beneficiary use of a low-value service. The composite measure variance is then calculated as

$$\widehat{\sigma^{2(e)}} = \sum_i \left(\sum_k w_k \varepsilon_{ijk}\right)^2 \cdot \frac{n_j}{n_j - 1}$$

where n_j is the number of beneficiaries assigned to the organization during the study period. This formula can be derived via Taylor series approximation of $\widehat{R_j}$.

We estimated the 95% confidence interval for the across-organization standard deviation by bootstrapping. Specifically, we obtained 1,000 parameter estimates by repeatedly drawing observations from the set of organizations with replacement and running the Fay–Herriot model. Noting that these parameter estimates had a roughly normal distribution, we used a normal approximation, calculating the 95% confidence intervals as the parameter point estimate plus or minus 1.96 times the standard deviation of bootstrapped parameter estimates.

We also used a normal approximation to calculated the 90th/10th percentile ratio, another measure of organizational variation. This was calculated based on a normal distribution centered at the unadjusted mean number of low-value services among all organizations (Table 1), with a standard deviation estimated via the Fay–Herriot model. Specifically, the 90th percentile was calculated as the grand mean plus 1.28 times the adjusted standard deviation, and the 10th percentile was calculated as the grand mean minus 1.28 times the adjusted standard deviation.

Consistency

The parameters of interest in our consistency analysis were the pairwise correlations between each service categories' composite scores. The methods for producing these estimates were extremely similar to those of the variation analysis. The same case-mix adjustment models that were used for the variation analysis were used for the consistency analysis. The distinguishing feature of the consistency analysis was that we calculated multiple composite scores, one for each of the six clinical categories of low-value services. In the variation analysis, the single composite measure was constructed from all 31 component services. In the consistency analysis, the six composite measures were each constructed from only the *K* component services that fall within the same clinical category, *c*. Thus, the formula for composite scores is:

$$\widehat{R_{cJ}} = \sum_{k=1}^{K_c} w_k \hat{r}_{jk}$$

Similarly, the estimated variance of each composite measure was a function of the prediction errors for the *K* services:

$$V_{cj} = VAR(\widehat{R_{cJ}}|R_{cJ}) = \sum_{i} \left(\sum_{k=1}^{K_c} w_k \varepsilon_{ijk} \right)^2 \cdot \frac{n_j}{n_j - 1}$$

Because there are six composite measures per organization, the corresponding Fay–Herriot model is multivariate normal rather than univariate. Specifically, the model is:

$$\widehat{R_J} = \begin{pmatrix} \widehat{R_{1J}} \\ \vdots \\ \widehat{R_{6J}} \end{pmatrix} \sim N(R_j, V_j), \text{ where}$$

$$R_j = \begin{pmatrix} R_{1j} \\ \vdots \\ R_{6j} \end{pmatrix} \sim N(\mu, \Sigma) \text{ and } V_j = diag(V_{1j} \dots V_{6j})$$

This model includes correlated organizational random effects, and the correlation between each pair of service domains is extracted from Σ. These estimated correlations between random effects are presented in Table 4. Again, 95% confidence intervals were estimated via bootstrapping, with the normal approximation described above.

Persistence

The parameter of interest for our persistence analysis was the correlation between organizations' low-value service composite scores in 2010 and 2011. Like the variation analysis, the persistence analysis employed composite measures that included all 31 low-value services. Like the consistency analysis, the persistence analysis involved constructing multiple composite scores, one for 2010 and one for 2011, which would be included in a multivariate correlated random effects model. One distinctive obstacle for estimating organizational persistence in behavior is the problem of autocorrelation in beneficiary outcomes over time. This problem did not arise for the variation and consistency analyses, since those composite measures were based on averaging ϵ_{ijkt} across all years of the study period. Organizational behavior could artificially appear correlated over time if positive temporal autocorrelation in a patient's service use were driving the result.

We purge our samples of this potential autocorrelation by estimating 2010 and 2011 composite measures for mutually exclusive sets of beneficiaries. Because

no beneficiaries are present in both of the modified 2010 and 2011 samples, autocorrelation in patient outcomes can no longer introduce bias into the correlation estimate. In order to maintain a representative sample of beneficiaries, we randomly assign each beneficiary who is present in our 2010 and/or 2011 sample to be included in either the 2010 sample (50% chance) or 2011 sample (50% chance). Then, we drop all beneficiary observations that do not occur in the assigned year. For instance, if a bene appeared in 2010 and in 2011, and is assigned to 2010, only their 2010 observation will be included. If a bene appeared only in 2011 and was assigned to 2010, then they will not appear in the final sample. Note that many more beneficiary observations would have been dropped if we used more than two years of data in our persistence analysis.

Following these modifications to our sample, we repeated the case-mix adjustment regressions using the new 2010-2011 sample. Component measures $\hat{r}_{jkt}$ were then calculated at the organization-year level as the average of ϵ_{ijkt} within each organization for 2010 and 2011. Similarly, The composite measure $\widehat{R_{Jt}}$ is constructed from all 31 measure components as $\widehat{R_{Jt}} = \sum_k w_k \hat{r}_{jkt}$. Since all beneficiaries are only present in the data for a single year, the previously described prediction error term is now calculated as:

$$\varepsilon_{ijkt} = \frac{\epsilon_{ijkt} - \hat{r}_{jkt}}{x_{ijk}}$$

The variance of the composite measure estimates are each calculated as

$$V_{Jt} = \widehat{VAR(\widehat{R_{Jt}}|R_{Jt})} = \left(\sum_i \left(\sum_k w_k \varepsilon_{ijkt}\right)^2\right) \cdot \frac{n_j}{n_j - 1}$$

where n_j equals the total number of beneficiaries in the modified sample who were assigned to each organization in both 2010 and 2011. To estimate the correlation in organizational behavior over time, we fit the following model:

$$\widehat{R_J} = \begin{pmatrix} \widehat{R_{J2010}} \\ \widehat{R_{J2011}} \end{pmatrix} \sim N(R_j, V_j), \text{ where}$$

$$R_j = \begin{pmatrix} R_{j2010} \\ R_{j2011} \end{pmatrix} \sim N(\mu, \Sigma) \text{ and } V_j = diag(V_{j2010}, V_{j2011})$$

Again, correlations between the random effects are extracted from Σ. These correlations are presented in Table 3.

Pioneer ACO Evaluation

Prices

Constructing standardized prices for each service allowed us to categorize services by price (high-price vs low-price) and to include a dollar-denominated measure of low-value service use as a study outcome. For each measure, a standardized price was calculated as the median of total allowed charges (from Medicare, beneficiaries and other payers) for relevant services in a care episode. Prices were calculated based on services detected in the first year of our study period.

For 25 of 31 measures, relevant services consisted of the detected service and other specific services delivered on the same day. For example, venipuncture is included as a relevant service for PSA screening. For the remaining six measures, which detected procedural/surgical services, it was not possible to comprehensively specify

the many CPT codes that could be relevant to the service episode. As described in prior methods,[1] we employed alternate pricing methods for these measures based on total daily charges and/or inpatient prospective payments. For services sometimes performed in the outpatient setting (vertebroplasty, renal artery angioplasty, arthroscopic knee surgery, and spinal injections), price was estimated based on the sum of Carrier and Outpatient charges during the day of the procedure. For surgical procedures occurring near-exclusively in the inpatient setting (carotid endarterectomy and PCI), price was estimated based on the sum of allowed Carrier charges during the procedure date and the spending allowed by the MS-DRG in the MedPAR file. CPT and MS-DRG codes for relevant services are included in column 3 of Table A2.1.

In order to ensure that prices were consistent across measures, prices for identical services included in multiple different measures (e.g. head imaging for syncope and head imaging for headache) were based on a pooled set of care episodes detected by both measures. For measures that include multiple services with substantial variation in price, we calculated a standardized price for each service. For example, separate prices were calculated for stress testing involving only exercise treadmill testing and for tests including advanced imaging.

The 16 measures with the highest standardized prices were designated as high-price and the remaining 15 were designated as low-price. Each beneficiary's annual spending on detected services was calculated by multiplying his or her annual count of each service by its standardized price.

Qualifying Indicators

Many beneficiaries do not fit the demographic or clinical characteristics needed to qualify for potential receipt of services we measured. Failure to account for differential changes in the qualifying characteristics of beneficiaries in the ACO vs non-ACO groups could introduce bias into our difference-in-difference estimations. As a result, we include binary indicators of measure qualification as covariates in our models. These qualification criteria are included in Table 2.1.1. We avoided qualification criteria based on symptoms (i.e. back pain or headache) since whether a beneficiary meets such criteria could be influenced by changing provider practice patterns. Based on these criteria, fifteen binary indicators for measure qualification were constructed (some applying to multiple measures) for each beneficiary in each year of our study sample.

We conducted a sensitivity test in which qualifying indicators were omitted from regressions estimating differential changes in the count of low-value services and associated spending. Results were extremely close to those presented in Table 2.2.3. In these analyses, the start of Pioneer contracts was associated with a differential reduction of 0.8 low value services per 100 beneficiaries in the ACO group ($P<0.001$) and a differential reduction in spending on these services of $455 per 100 beneficiaries ($P=0.005$). These corresponded to reductions of 2.0% and 4.4%, respectively.

Baseline Outcomes and Mean Reversion

Two measures were constructed to assess organizations' baseline rate of low-value service delivery in 2008. The first, a measure of service area rates, isolated practice patterns that can be attributed to geography. The measure is based on the risk-

adjusted count of low-value services in each ACO's geographic service area among beneficiaries in the non-ACO group. This parameter is calculated by performing linear regression of the total counts of low-value services on a set of HRR indicators as well as the demographic and clinical controls appearing in our main analyses, for the 2008 non-ACO sample. Then, for each ACO, we calculate an average of the HRR coefficients that is weighted by number of the ACO's beneficiaries in each HRR. ACOs are then categorized as serving areas with high or low levels of low-value services according to whether the weighted average falls above or below that of the median ACO. The second measure assesses the ACO's baseline performance relative to its geographic service area. This measure is calculated using the full ACO and non-ACO sample by regressing low-value service counts on beneficiary covariates, HRR indicators and ACO indicators. ACOs are classified as deviating above or below the HRR average based on whether their fixed effects coefficients are greater than zero or less than zero.

These measures, based on 2008 data, are predictive of ACO characteristics in the 2009-2011 pre-contract period. In the 2009-2011 ACO group, the adjusted utilization of low-value services relative to the local mean was 5.5 services per 100 beneficiaries higher for ACOs with levels greater than the local mean in 2008 that that of ACOs with levels lower than the local mean in 2008 ($P<0.001$). Also, in 2009-2011, the adjusted count of low-value services in ACO service areas was 12.2 services per 100 beneficiaries higher for ACOs classified as high use services areas in 2008 than for those classified as low use service areas ($P<0.001$).

In order to minimize the possibility of bias from regression to the mean, these baseline characteristics were measured in 2008, before the start of the study period. Bias from regression to the mean may occur whenever analyzing whether high or low

baseline levels of an outcome predict future changes in that outcome. This possibility is unlikely in our study, however, since there was no evidence of regression to the mean during the pre-contract period. Indeed, ACOs with high baseline utilization levels relative to their service area saw adjusted low-value service utilization grow somewhat faster by 0.5 services per year during 2009-2011 (P=.06), a temporal trend in the opposite direction as would be predicted by regression to the mean.

Analyses Adjusting for Pre-Contract Trends

We repeated our main analyses with models that test and adjust for the presence of non-parallel trends in outcomes between the ACO and non-ACO groups during the pre-contract period. These models were of the following form:

$$E(Y_{i,t,k,h}) = \beta_0 + \beta_1 ACO_{indicators_k} + \beta_2 HRR\_indicators_h \times Year_t + \beta_3 ACO\_Group_k \times 2012_t + \beta_4 ACO\_Group_k \times Year\_Continuous_t + \beta_5 Covariates_{it}$$

where "Year_Continuous" is the year of study observation, specified continuously (2009-2012). This model differs from those described in the body of the manuscript because of the inclusion of the "ACO_Group×Year_Continuous" term, whose β_4 coefficient represents the difference in linear annual trend between the ACO and non-ACO group in the pre-contract period. The magnitude and statistical significance of this coefficient serve as our test for non-parallel trends in the pre-contract period. There was no statistically significant evidence of non-parallel trends for any of the outcomes reported in Table 3. For example, during the pre-contract period, the adjusted annual count of low-value services in the ACO group changed at a rate of 0.1 services per 100 beneficiaries per year faster than the non-ACO group (P=0.74), and adjusted spending

on low-value services in the ACO group changed at a rate of $20 per 100 beneficiaries per year slower (P=0.88).

β_3 remains the coefficient representing the estimated effect of Pioneer contracts in 2012. However, this estimate now reflects the assumption that, in the absence of the Pioneer contract, outcomes in the ACO group would have continued according to the estimated prior linear trend, which may not have been parallel to that of non-ACO beneficiaries in the region. This may not be a reasonable assumption, especially if a pre-contract divergence in trends is due to randomness. Such divergence in pre-contract trend would tend to be followed by convergence in the post-contract period rather than continued divergence, due to regression to the mean. Importantly, introducing the trend term into this model increases the confidence intervals on the β_3 coefficient because the estimates now incorporate the additional uncertainty with which these extrapolated trends were estimated. Still, we believe that these models may serve a useful purpose as a robustness test even though no statistically significant divergent trends were found in the pre-contract period. Trend-adjusted differential changes in low-value service frequency are presented in Table A2.2. The magnitudes of these estimates are largely similar to those presented in Table 2.2.3. Following trend-adjustment, the magnitude of the differential reduction in the count of low-value services was largely unchanged, moving from 1.9% to 2.1%, as was the magnitude of the differential decrease in spending on low-value services, moving from 4.5% to 4.1%.

In order to adjust our analyses of organizational subgroups for pre-contract trends, we introduced interactions between the organizational characteristics of interest and the β_3 and β_4 terms. Results from these analyses are presented in Figure A2.1. Following trend adjustment, the estimated effects of Pioneer contracts were still

greater for ACOs with higher baseline levels of low-value services than their service area (-1.8 services per 100 beneficiaries) than for ACOs with lower baseline rates (-0.1 services per 100 beneficiaries, P=0.002 for difference), and there were still no statistically significant associations between ACO performance and other characteristics of the organizations.

Additional References

1. Schwartz AL, Landon BE, Elshaug AG, Chernew ME, McWilliams JM. Measuring low-value care in Medicare. *JAMA Internal Medicine* 2014; 174(7):1067–76.
2. Fay R, Herriot R. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 1979; 74(366):269–77.
3. Zaslavsky AM. Using hierarchical models to attribute sources of variation in consumer assessments of health care. *Statistics in Medicine* 2007; 26(8):1885–900.
4. O'Malley AJ, Zaslavsky AM. Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse. *Journal of the American Statistical Association* 2008;103(484):1405–18.
5. Agency for Healthcare Research and Quality. Instructions for Analyzing Data from CAHPS Surveys. 2015. Available from: https://cahps.ahrq.gov/surveys-guidance/survey4.0-docs/2015-Instructions-for-Analyzing-Data-from-CAHPS-Surveys.pdf
6. Wiener RS, Welch HG. Trends in the use of the pulmonary artery catheter in the United States, 1993-2004. *JAMA* 2007 ;298(4):423–9.

Table A2.2 Trend-Adjusted Differential Changes in Low-Value Service Frequency, by Service

Annual Count or Spending per 100 Beneficiaries	Trend-Adjusted Differential Change	95% CI	Trend-Adjusted Differential Change as percent of ACO Mean[a]	95% CI	P-Value
Total low-value services, no.	-0.9	(-1.5, -0.2)	-2.1	(-3.7, -0.5)	0.01
Total low-value service spending, $	-420	(-937, 98)	-4.1	(-9.1, 1)	0.11
Low-value services by clinical category, no.[b]					
Cancer screening	-0.2	(-0.5, 0.2)	-1.5	(-4.6, 1.6)	0.33
Testing	-0.3	(-0.6, 0.1)	-3.0	(-7, 1.1)	0.15
Preoperative Services	-0.1	(-0.2, 0.1)	-2.4	(-8.7, 3.9)	0.46
Imaging	-0.2	(-0.6, 0.2)	-1.4	(-4, 1.1)	0.26
Cardiovascular Tests and Procedures	-0.1	(-0.1, 0)	-7.5	(-14.8, -0.2)	0.04
Other Invasive Procedures	-0.1	(-0.4, 0.2)	-2.7	(-9.1, 3.7)	0.41
Low-value services by price, no.[b]					
High price	-0.2	(-0.7, 0.2)	-1.4	(-4.2, 1.4)	0.31
Low price	-0.6	(-1.2, -0.1)	-2.5	(-4.6, -0.4)	0.02
Low-value services by sensitivity to patient preferences, no[b]					
More sensitive	-0.7	(-1.4, -0.1)	-2.7	(-5, -0.3)	0.03
Less sensitive	-0.1	(-0.5, 0.2)	-0.9	(-3.6, 1.8)	0.51

ACO = Accountable Care Organization

[a] Calculated as the differential change divided by the adjusted 2012 mean for ACO group.

[b] Note that the sum of differential changes within each set of service categories equals the total differential change.

FIGURE. A2.1 Trend-Adjusted Differential Changes in Use of Low-Value Services in ACO vs. Control Group, by Baseline Use