This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL REPORT

Provider-Level Risk-Adjusted Quality Measurement for Inpatient Rehabilitation Facilities

Andrew W. Dick, Peter J. Huckfeldt, Hangsheng Liu, Hao Yu, Ateev Mehrotra, Susan L. Lovejoy, J. Scott Ashwood

Sponsored by the Medicare Payment Advisory Commission

HEALTH

The research described in this report was sponsored by the Medicare Payment Advisory Commission and was conducted within RAND Health, a unit of the RAND Corporation.

Published 2012 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

PREFACE

Quality metrics play an increasingly important role in the evaluation and reimbursement of post-acute providers. Currently, it is difficult to ascertain whether changes in inpatient rehabilitation facility (IRF) patient outcomes are due to changes in treatment or the case mix of patients seen in IRFs. To address this issue, the Medicare Payment Advisory Commission (MedPAC) contracted with the RAND Corporation (1) to develop risk-adjusted quality metrics at the provider level for IRFs, (2) to develop methods to address low case volume and uncommon events, and (3) to use those metrics to estimate national trends in IRF quality from 2004 to 2009. This report presents the results for five IRF outcomes: (1) functional gain, (2) discharge to the community, (3) 30-day readmission to acute care given discharge to the community, (4) 30-day readmission to skilled nursing facility (SNF) given discharge to the community, and (5) discharge directly to acute care. The risk-adjustment models presented here minimize the potential for selection and can ultimately be used as the basis for public reporting and quality-based reimbursement.

This report should be of interest to researchers involved either in inpatient rehabilitation facilities or the development of IRF quality reports or in outcomes studies that require risk adjustment. The study should also interest funders of IRF services, those working on risk adjustment methods in health services research, and patients (and their families) who have a need for IRF services.

CONTENTS

FIGURES

TABLES

SUMMARY

Quality metrics play an increasingly important role in the evaluation and reimbursement of post-acute providers, and inpatient rehabilitation facilities (IRFs) in particular. The Medicare Payment Advisory Commission (MedPAC) uses aggregate trends in quality measures, such as functional gain, to assess the adequacy of payments to providers. In addition, the Patient Protection and Affordable Care Act (ACA) requires IRFs to publicly report quality data or be penalized in annual payment updates. Eventually, quality data could be used in a value-based purchasing scheme for IRFs as is being developed for other post-acute settings under ACA.

In this report, we describe our development of risk-adjusted quality metrics for IRFs, focusing on five patient outcomes: (1) functional gain, (2) discharge to the community, (3) 30-day readmission to acute care, (4) 30-day readmission to a skilled nursing facility (SNF), and (5) discharge directly to acute care. We start by describing the development of our model, including our choice of independent variables, the selection of our model specification, and the identification of a minimum patient volume to support a quality metric. Next, we present results from our risk-adjustment estimation. We identify changes in model parameter estimates across years in the sample period and decompose average patient outcome trends into changes stemming from the case-mix composition of patients seen in IRFs and those due to changes in expected outcomes given the observable case mix. Finally, we examine the persistence of quality estimates and differential trends in quality by IRF provider characteristics.

For each outcome, we specified a model that contained individual socioeconomic and demographic characteristics, comorbid condition indicators, Impairment Group Code (IGC) indicators, and age-by-sex interactions. We estimated nine alternative specifications for each model and chose the model with the best fit. In each case, the best model was either the full model or the model that dropped the age-by-sex interactions. Model fit was good, but there remained a substantial amount of unexplained outcome variation by IRF.

We considered two main modeling approaches for the IRF risk adjustment: (1) random-effects or hierarchical models and (2) fixed-effects models in which each IRF's contribution to quality is determined by a fixed indicator in the model. For each of the outcomes models, we found strong evidence to reject the random-effects assumption, indicating that quality measures based on random-effects estimates would be biased. We therefore adopted the fixed-effects approach. In the case of this application, even though the number of IRFs was large (1,400), the very large data set (2.5 million observations) made it feasible to estimate fixed-effects models. One notable advantage of the fixed-effects approach is that the models generate an empirical representation of the distribution of quality across IRFs. Unlike random-effects models, which assume a distribution for IRF quality, the fixed-effects approach imposes no structure. In addition, by manipulating the models so that each IRF's estimate represents its difference from the average-quality IRF, the quality metrics are relatively easy to manipulate and interpret.

Although our study sample contained approximately 2.5 million observations, not every IRF had large patient volume. Thus, we developed criteria for determining if the quality estimate for an IRF was too uncertain or unreliable to merit reporting. Starting with the assumption that the estimated uncertainty (standard error or confidence interval of the quality measure) is the best means for determining the measure's reliability, we examined the relationship between IRF volume and the quality measure's uncertainty, and we examined various criteria that used only IRF volume to determine the omission of IRFs. We found that the use of a volume standard resulted in either the omission of too many IRFs (many of which had reliable estimates of quality) or the inclusion of IRFs that had very uncertain quality estimates. We therefore recommend the use of a combined standard that requires a minimum patient volume (we used n = 30) to eliminate very-small volume IRFs *and* a maximum uncertainty level (confidence interval > one standard deviation of the outcome rate across institutions) in order to exclude IRFs for which the quality estimates were not precise. We found that these criteria together resulted in the exclusion of a very small number of IRFs.

We examined model stability by comparing model estimates based on data from early in the study period with model estimates based on data from late in the study period. In

every case but one (30-day readmission to SNF comparing 2006–7 estimates with 2008–9 estimates), we found evidence that the model parameter estimates changed over time. This implies that caution should be used when making model estimates using data pooled over several years. It also suggests that risk-adjustment models will need to be recalibrated regularly. As a result, we estimated year-specific risk-adjustment models to generate year-specific quality indicators for each IRF. These indicators can be compared either within each year, to reveal any IRF's quality relative to other IRFs or to the average IRF, or over time to reveal trends in quality.

We also used the models to examine trends in quality during the study period. A comparison of raw and adjusted outcome rates shows the contribution of real quality changes and case-mix changes, revealing that real improvements in quality have been masked by increasing severity in case mix during each year of the study and for each of the outcomes considered. Among the key findings are the following:

1. *The raw outcome rates for four of the five quality measures worsened over the study period*. Functional Independence Measure (FIM) gain improved by about two points, but discharge to the community (18 percent to 11 percent), 30-day readmission to acute care (10.5 percent to 11.5 percent), 30-day readmission to SNF (3.1 percent to 3.7 percent), and discharge directly to acute care (8.6 percent to 10.25 percent) all worsened.
2. *The declining raw rates were caused by worsening case mix during the period*. After adjusting for case mix, we found that quality improved on every metric; in the case of FIM gain, quality improved by more than the improvement in the raw rate.
3. *Quality at individual IRFs persisted over the study period*. We identified two cohorts of IRFs using 2004 quality estimates: the lowest-quartile performers and the highest-quartile performers. We then examined the average quality of those cohorts in each of the study years, and we found that the 2004 high-performing IRFs continued to be high performers throughout, that the 2004 low performers continued to be low performers, and that the gap in quality did not narrow substantially.

Our work indicates that the existing data provide ample information to estimate risk-adjustment models for the purpose of examining and reporting IRF quality. Furthermore, our findings show that, because of the substantial worsening of case mix, the use of high-quality risk-adjustment models as the basis for quality reporting is essential for revealing overall quality trends in the market. Finally, we find that the risk-adjustment parameter estimates change over time, indicating that the models should be recalibrated regularly and, if feasible, year-specific models should be used.

ACKNOWLEDGMENTS

We would like to acknowledge the assistance and guidance of our MedPAC project officer, Christine Aguiar, policy analyst. We also thank the RAND quality assurance peer reviewers, John Engberg and David Auerbach, for their very helpful comments on the report.

ABBREVIATIONS

ACA	Patient Protection and Affordable Care Act
AIC	Akaike Information Criterion
AMI	acute myocardial infarction
BIC	Bayesian Information Criterion
CI	confidence interval
CMS	Center for Medicare and Medicaid Services
FIM	functional independence measure
IGC	Impairment Group Codes
IRF	inpatient rehabilitation facility
MedPAC	Medicare Payment Advisory Commission
MedPAR	Medicare Provider Analysis and Review
OR	odds ratio
PAI	Patient Assessment Instrument
RIC	Rehabilitation Impairment Category
SNF	skilled nursing facility

1. INTRODUCTION

Quality metrics play an increasingly important role in the evaluation and reimbursement of post-acute providers, and inpatient rehabilitation facilities (IRFs) in particular. The Medicare Payment Advisory Commission (MedPAC) uses aggregate trends in quality measures, such as functional gain, to assess the adequacy of payments to providers. In addition, the Patient Protection and Affordable Care Act (ACA) requires IRFs to publicly report quality data or be penalized in annual payment updates. Eventually, quality data could be used in a value-based purchasing scheme for IRFs, as is being developed for other post-acute settings under ACA.

In order to distinguish IRFs from acute care hospitals, the Center for Medicare and Medicaid Services (CMS) requires that 60 percent of a facility's caseload be diagnosed with conditions deemed to require intensive inpatient rehabilitation. Starting in 2004, this list of conditions was altered, most notably to exclude many types of joint replacement. The effect of this policy change was a lower volume but higher severity of IRF patients (MedPAC, 2011). The change in composition of IRF patients over time makes the interpretation of quality measures more challenging. For example, MedPAC (2011) shows that functional gain (as measured by Functional Independence Measure (FIM)) has increased between 2004 and 2010. However, it is difficult to ascertain whether these improvements in outcomes are due to improved treatment in IRFs or to the fact that the types of patients admitted to IRFs had a greater potential for functional gain, as noted in MedPAC (2011).

Risk adjustment has the potential to improve the comparability of quality metrics both across providers and over time. Risk-adjusted quality measures can be used internally by MedPAC to determine payment adequacy and externally for public reporting. Eventually, they could form the basis of a value-based purchasing scheme for IRFs. Such models allow policymakers to distinguish between aggregate changes in quality measures due to treatment patterns and those due to the composition of patients seen in IRFs.

In this report, we describe our development of risk-adjusted quality metrics for IRFs, focusing on five patient outcomes: (1) functional gain, (2) discharge to the community, (3) 30-day readmission to acute care, (4) 30-day readmission to a skilled nursing facility (SNF), and (5)

discharge directly to acute care. We start by describing the development of our model, including our choice of independent variables, the selection of our model specification, and the identification of a minimum patient volume to support a quality metric. Next, we present results from our risk-adjustment estimation. We identify changes in model parameter estimates across years in the sample period and decompose average patient outcome trends into changes stemming from the composition of patients seen in IRFs and those due to changes in expected outcomes as a function of observable characteristics. Finally, we examine the persistence of quality estimates and differential trends in quality by IRF provider characteristics.

2. METHODS

Data and Choice of Risk Adjustors

Our primary data source is the Medicare Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF PAI) from 2004 to 2009. At admission, patients' personal and demographic characteristics, admission class, preadmission living status, insurance information, impairment group (i.e., reason for rehabilitation), etiologic diagnosis (i.e., diagnosis leading to impairment group), up to ten comorbid conditions and preexisting complications, functional status, cognitive status, and other information are entered on the IRF-PAI. The impairment requiring rehabilitation, cognitive and functional status at admission, age, and the severity of comorbidities are then used to assign each patient to a case mix group, which determines the amount of the prospective payment a facility will receive (MedPAC, 2010). Cognitive status, functional status, and new complications are also recorded at discharge in order to measure quality of care.

The quality outcomes used in our models are derived from the IRF PAI data. The first outcome is improvement in functional status, as measured by the change in the FIM scale between admission and discharge. The FIM scale contains 18 items, 13 of which are physical or motor items and 5 of which are cognitive items. Each item is scored on a seven-point scale, where 1 indicates low function and 7 indicates high function. The total FIM score ranges from 18 to 126; the FIM motor score ranges from 13 to 91. The IRF PAI also indicates whether the patient has been discharged back to an acute care hospital.

We augmented IRF PAI data with the Medicare Provider Analysis and Review (MedPAR) acute claims occurring within a year prior to each IRF stay to capture admissions to a SNF or an acute care hospital within 30 days of discharge from the IRF. We recorded all comorbidities in acute claims from the prior year, as well as complications from the preceding acute stay. We also linked IRF patients to the Medicare Denominator File to obtain a more complete set of demographic characteristics. Finally, we linked the provider information from the Medicare Provider of Services File in each year, including freestanding versus hospital-based status, urban and rural location, facility ownership status (i.e., for-profit, government-run, or nonprofit), and bed

size. We used provider characteristics to examine quality trends separately by provider characteristics, not in the risk adjustment.

We selected patient characteristics that are likely to influence post-acute outcomes as risk adjustors for our models. Demographic characteristics include patient age (indicator variables for five-year intervals), race and ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, and other race), marriage status, dual eligibility (for Medicare and Medicaid), and disability status. We included comorbidities identified by Elixhauser (1998) that are predictive of hospital charges, length of stay, and patient health outcomes.[1] We controlled for complications present at admission from Iezzoni et al. (1994) that are likely to have a continued effect after hospital discharge and could influence post-acute outcomes—in particular, dropping transient metabolic derangements and medication side effects. We also included additional complications that may affect post-acute outcomes in a Medicare population. Our final list included postoperative pulmonary compromise; post-operative gastrointestinal hemorrhage; cellulitis or decubitus ulcer; septicemia; pneumonia; mechanical complications due to a device, implant, or graft; shock or arrest in the hospital; postoperative acute myocardial infarction (AMI); postoperative cardiac abnormalities other than AMI; venous thrombosis and pulmonary embolism; procedure-related perforation or laceration; acute renal failure; delirium; and miscellaneous complications.

We also strove to capture changes in patient composition attributable to changes in the 75 percent rule implemented starting in 2004. Prior to the revision of the 75 percent rule in April 2004, a facility was designated an IRF (and thus exempt from the acute Prospective Payment System) if at least 75 percent of its patients in a facility received treatment for the following conditions: stroke, spinal cord injury, congenital deformity, amputation, major multiple trauma, fracture of femur, brain injury, polyarthritis, neurological disorders, or burns (Carter et al., 2003). However, prior to 2004, the 75 percent rule was not enforced, and polyarthritis was often interpreted to include lower-extremity joint replacement. The new regulations implemented in

[1]These comorbidities include congestive heart failure, valvular disease, pulmonary circulation disorders, peripheral vascular disorders, hypertension, paralysis, other neurological disorders, chronic pulmonary disease, diabetes (uncomplicated), diabetes (complicated), hypothyroidism, renal failure, liver disease, peptic ulcer disease excluding bleeding, AIDS, lymphoma, metastatic cancer, rheumatoid arthritis/ collagen vascular diseases, coagulopathy, obesity, weight loss, fluid and electrolytes disorders, blood loss anemia, deficiency anemias, alcohol abuse, drug abuse, psychoses, and depression.

2004 mandated enforcement of the 75 percent rule, but also replaced the polyarthritis category with a more specific list that includes three arthritis conditions for which appropriate, aggressive, and sustained outpatient therapy has failed, and joint replacement for both knees or hips when surgery immediately precedes admission, when BMI is greater than or equal to 50, or when age is greater than 85 (MedPAC, 2010). Enforcement of the 75 percent rule was originally to be phased in over a five-year period, starting from 50 percent in 2004. However, it was decided in 2007 that the compliance threshold would stay at 60 percent (MedPAC, 2010).

With insufficiently specific patient controls, differences in severity or composition within patient categories across facilities could be misattributed to facility quality after regulatory changes. Adequately controlling for the impairment requiring rehabilitation is an important dimension of capturing changes in composition. Patients are categorized at admission into one of 85 Impairment Group Codes (IGCs), which are then nested in 21 Rehabilitation Impairment Categories (RICs). Regulatory changes in 2004 map more closely to IGCs than RICs. Lower-extremity joint replacement patients count toward the 75 percent rule if the replacement is bilateral under certain circumstances. Thus, while the overall Replacement of Lower Extremity Joint RIC is too broad to capture this change, the specific IGC codes within this RIC include separate categories for unilateral versus bilateral replacement for hip, knee, and joint. Paired with the Elixhauser obesity control and age controls, the patient characteristics included in our risk-adjustment models map to these regulatory changes. The Arthritis IGCs, in contrast, are not extensively disaggregated and cannot be directly mapped to the changes in arthritis categories counting toward the threshold. However, in general, linearly controlling for IGC and the extensive list of patient demographic characteristics, comorbidities, and complications provides an approximate mapping to the regulation changes and captures most changes in patient composition that could bias quality estimates.

Specification Tests

Our basic model specifies patient outcomes as follows:

$$y_{ipt} = \alpha_p + \beta x_{ipt} + u_{ipt} \quad (1)$$

where y is an outcome for patient i in facility p at time t, α is a facility-specific effect, x includes demographic and health characteristics that influence patient outcomes, and u is a person-level error term. Note that all health-related measures that are used for controls must be defined at or prior to admission to the IRF to avoid the possible influence of IRF quality on the measures. The facility-specific effect, α, can be modeled using either random-effects or fixed-effects models. A random-effects specification assumes that the facility-specific effect is distributed independently of patient characteristics (x). Random-effects models estimate model parameters using variation both across and within facilities, leading to more-efficient estimation. However, the assumption of independence of α and x is not necessarily valid. Fixed-effects models make only the assumption that the error term (u) is uncorrelated with patient characteristics (x), and they impose no restrictions on the distribution of quality across IRFs. In addition, the estimated fixed effects themselves form an intuitive quality metric. However, we were unable to estimate fixed-effects models with too few observations per panel member, and models with many panel members are computationally challenging to estimate. Our analyses dataset is very large, mitigating the former issue. Improvements in technology (e.g., dual core processors and statistical software harnessing this computational power) make it feasible to estimate these models.

We tested the random effects assumption of independence between the facility-specific effect and patient characteristics using a Hausman test. Under the null hypothesis of independence, the random-effects and fixed-effects parameter estimates for the x variables are both consistent and thus should be equivalent. The Hausman test uses the following test statistic:

$$\mathrm{H} = (\beta_{RE} - \beta_{FE})'\{V(\beta_{RE} - \beta_{FE})^{-1}\}(\beta_{RE} - \beta_{FE})' \qquad (2)$$

where $(\beta_{RE} - \beta_{FE})'$ represents the difference between the random-effects (RE) and fixed-effects (FE) estimates of parameter estimates, and V is the variance. H is distributed with a chi-squared distribution; large values of H imply statistically significant differences between the fixed-effects and random-effects estimates and lead us to reject the null hypothesis and, therefore, the random-effects formulation.

We outline the results of Hausman tests in later sections but note here that we found consistent and strong evidence to reject the equivalence of the random-effects and fixed-effects parameter

estimates. This implies the correlation of provider effects and patient characteristics, and thus we used fixed-effects specifications in our risk-adjustment models. FIM gain is a continuous variable; as such, we estimated linear fixed-effects regressions in the risk-adjustment models. The other outcomes (discharge to community, readmission to acute hospital, readmission to SNF or long-term care, and discharge directly to acute care) are dichotomous—equal to one or zero—and thus we estimated logistic regressions. A potential drawback of fixed-effects logistic regression models is that each panel member (e.g. IRF) must have successes and failures to be included in the estimation. Thus, these models are sometimes referred to as conditional fixed-effects models. In addition, because the sample selection is conditional on the dependent variable, the fixed-effects estimates may be biased if the number of cases within panel members (IRFs) is small. The very large sample sizes per facility available for our analyses mitigate this concern, but they do create an additional reason to be concerned with quality estimates for institutions that have few cases.

For all outcomes, we estimated facility fixed effects directly for use as the quality metric by including indicator variables for each facility. The model is specified such that the reference (or omitted) category is the average risk-adjusted outcome across facilities, not weighting by patient volume. For example, in the case of FIM gain, the quality metric for a given facility can be interpreted as the average FIM gain in the facility relative to the IRF average FIM gain, holding constant patient demographic and health characteristics. The fixed-effects coefficient estimates for the FIM gain model are directly interpretable as the increment in outcome due to the quality of the facilities. For the dichotomous outcomes, the coefficient estimates are difficult to interpret directly. Therefore, rather than using the fixed-effects coefficient estimates directly, we exponentiated them and multiplied them by the overall outcome rate in the sample, so that the quality metrics are directly interpretable in units of the outcome.

We evaluated an extensive set of models using varying combinations and functional forms for patient characteristics. To evaluate alternative specifications, we created estimation and validation samples from the 2004 through 2009 sample period, where two-thirds of patients in each facility were randomly assigned to the estimation sample and one-third was assigned to the validation sample. We estimated each specification in the estimation sample, and then examined the fit of the model in the validation sample. For continuous variables (i.e., FIM gain and FIM gain–motor), we compared the mean square predicted error using the predicted values in the validation sample

across different specifications. For dichotomous variables, we calculated C-statistics in the validation sample. The C-statistic gives the fraction of total 0–1 outcome pairs in the data where the predicted outcomes for the person with outcome = 1 is higher than the predicted value for the person with outcome = 0. For all models, we calculated Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) goodness-of-fit measures in the estimation sample. These statistics penalize the model log likelihood for larger numbers of regressors; smaller statistics indicate preferable specifications.

Our first specification is a full model that includes the complete set of demographic variables, comorbidities, and complications; an indicator for admission from the community; an indicator for prior IRF stay; and FIM cognitive and motor scores at admission; and uses IGC to control for the reason for rehabilitation. In addition, we allowed for interaction effects between gender and age. Finally, we allowed for interaction effects for having multiple Elixhauser comorbidities in a system and for having Elixhauser comorbidities in more systems, where the systems are abuse (drug and alcohol), cancer/immune system, cardiac, blood diseases, mental health, neurologic, and weight disorders. We then estimated other specifications that omit certain controls or switch out controls for more parsimonious alternatives. For each outcome variable, we compared the goodness-of-fit statistics discussed above to select the best-performing model.

Minimum Cases to Support Quality Metric

IRF provider fixed effects are less precisely estimated as the number of patients seen in a facility decreases. We determined inclusion criteria based on the uncertainty of the estimated quality metric and a minimum sample size for an IRF such that only IRFs with valid and reliable quality metrics are considered. We proposed limiting estimates to providers for which the quality metrics confidence interval falls below some threshold. We experimented with setting this threshold as the estimated standard deviation of an outcome variable in the full sample or, alternatively, as the minimum clinically significant change estimated in the rehabilitation literature. We then identified the minimum patient volume at which the confidence interval is always smaller than the maximum threshold. Ultimately, we determined that using a standard based only on a volume threshold had problems. A volume threshold that eliminates all IRFs with unsuitable estimates (based on the estimated uncertainty) would result in the omission of a

substantial number of IRFs, many of which would have very reliable estimates. A lower volume threshold, however, would result in the inclusion of very uncertain quality estimates for a small set of IRFs. We therefore recommend the use of a mixed standard that includes a minimum volume (below which an IRF would be omitted) *and* a maximum level of uncertainty (above which an IRF would be omitted). We set the minimum volume threshold to be 30 observations and the maximum uncertainty threshold based on the confidence interval being no greater than one standard deviation of the outcome. This resulted in the smallest number of omissions while maintaining the reliability of the results. We also show that by pooling patients across years, many of the omitted low-volume facilities could be included in the reporting sample.

Stability of Model Parameters

After selecting the best-performing risk-adjustment specifications for each outcome, we examined the stability of risk-adjustment model parameters over time. First, we used a sample of IRF patients from 2004 and 2007–9 and estimated the risk-adjustment models, allowing the coefficient estimates to differ in two periods (2004 versus 2007–9). We then formed tests of the equivalence of the estimates in the two periods. Next, we used a sample of IRF patients from 2006–9 and estimated the model's parameters allowing the estimates from 2006–7 to differ from those of 2008–9. We then formed tests of the equivalence of the estimates in the two periods.

Changes in model parameters across years could reflect changes in the quality of care offered by IRFs. We investigated this further by decomposing unconditional trends in patient outcomes into changes in model parameters across years (potentially representing changes in practice patterns and quality), and those stemming from changes in patient composition. To look at changes driven by treatment patterns, we estimated risk-adjustment models for the IRF patient cohort in each year between 2004 and 2009, expressing quality as a function of patients' observable characteristics. We then applied each year's treatment function to a fixed cohort of IRF patients. As patient composition is held constant, changes in this measure across (model) years are driven entirely by changes in how patients' observable characteristics predict outcomes over time. Such shifts could represent treatment patterns or quality changes on the part of providers.

Next, we examined changes in patient outcomes driven by the composition of patients seen in IRFs. We estimated the risk-adjustment model for each outcome for each cohort of IRF patients.

We then projected quality outcomes in each IRF patient cohort across years, holding the treatment function fixed. Changes in average predicted outcomes across years in this case only reflect changes in the composition of patients; predicted outcomes as a function of observable characteristics are held fixed.

Contrasting the treatment pattern effects and composition effects will provide evidence on the relative contributions of treatment patterns and patient composition to unconditional trends in patient outcomes across IRFs.

Persistence in Quality

We generated matrixes with risk-adjusted quality metrics in each year for each facility (i.e., the facility fixed effects). From these data, we investigated the persistence of quality estimates by stratifying 2004 IRF providers into quartiles of quality and then examining average quality in subsequent years by 2004 quartile.

Choice of Quality Metric

As described in Chapter Three, we rejected random-effects models in favor of fixed-effects models. Thus, we adopt a quality measure that is derived from the fixed-effect coefficient estimates. Our basic model specifies patient outcomes as linear in X and IRF-specific fixed effects in the case of FIM gain, and as linear in the log of the odds ratio (OR) (logit models) in the case of the dichotomous dependent-variable models (see Eq. (1) above). In the linear case (FIM models), the estimated α vector represents the differences between the average IRF and each of the individual IRFs in units of the outcome. We interpret these directly as measures of Q. For the logistic regressions, coefficient estimates of dichotomous independent variables are difficult to interpret directly. They are in units of the standard deviation of the logistic distribution. Typically, estimates are presented as odds ratios because of the convenient characteristics of the logistic distribution function. The OR for one IRF relative to the referent IRF = $\exp(\alpha)$, where α is the estimated fixed-effect coefficient for the given IRF. We design the models so that the referent IRF is the average institution and has an OR of 1.

A standard method for defining quality (Q) in the literature has been to use the observed (O) to expected (E) ratio times the sample mean outcome rate (P0), where O is the observed rate for the institution and E is the expected rate for the institution. (O/E) x P0 is convenient because it

expresses the Q measure on the metric of the outcome. A comparison of an institution's Q to that of the average is interpreted as, for example, excess mortality at the institution.

We take a similar approach, also used in the literature, which is based on the notion that the OR is an approximation of O/E, particularly when the outcome incidence is low. We therefore define the Q as $\exp(\alpha) \times P0$. Because the referent OR is 1, Q for the referent institution is then the average outcome in the entire sample. Q is approximately in the metric of the outcome and, when compared to the sample mean rate, can be interpreted as excess outcome (e.g., excess percentage of 30-day readmissions to acute care). As P0 increases from very low values, the Q metric expands and is no longer in the metric of the outcome, but all of the important properties hold. Formally, the OR for facility 1 is defined as

$$OR = [P1 / (1 - P1)] / [P0 / (1 - P0)].$$

Therefore

$$Q = \exp(\alpha) \times P0 = OR \times P0 = [P1 / (1 - P1)] / [P0 / (1 - P0)] \times P0$$
$$= (P1 / P0) \times P0 \times (1-P0) / (1 - P1).$$

P1 is the IRF-specific rate for facility 1 and P0 is the expected rate, so (P1/P0) x P0 is O/E x P0. When the mean outcome rate (P0) is small, and deviations in outcomes for the facility of interest (P1) are small, the last ratio is close to 1. Thus, $Q = \exp(\alpha) \times P0$ is easily interpretable and convenient to calculate. When P is large, however, that last ratio can be large.

Although the above metrics are suitable for intra-year comparisons of relative quality, we form an additional metric (standardized IRF prediction) that eliminates the approximation error and facilitates comparisons over time even when the year-specific models have been estimated. To generate this metric, we assign each subject in a given year to each IRF (one at a time) and generate the mean (standardized) prediction for each IRF over the entire sample. In this way, we generate the expected outcome for a standardize distribution of case mix for each IRF in each year. Differences across IRFs within year are driven by the IRF fixed effects estimates. Difference both within and across IRFs over time, however, are driven both by the IRF fixed effects estimates and by the estimated case-mix parameters (which differ by year). Thus, the standardized

prediction metric accounts both for changes in risk-adjustment model estimates from year to year (for example, technical improvements in quality for a given type of individual) as well as the relative quality (within year) of each IRF, thereby facilitating comparisons over time.

3. RESULTS

Development of the Models

Choice of Independent Variables

We consulted the literature to define a set of base independent variables that are likely to influence rehabilitative success, disposition of discharge, and subsequent readmissions, and that are available in the IRF PAI, prior acute hospital claims, and the Medicare Denominator File. All health status related measures must be defined at the time of or before admission to the IRF so they are not influenced by the quality of the IRF. Variables from the Denominator Files include age (controlled using indicator variables for five-year intervals: under age 50, ages 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, 85–89, and 90 and above), race and ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, and other race), marriage status, dual eligibility status, and disability status. From the IRF PAI, we controlled for the primary reason for rehabilitation (using either RIC or IGC), whether a patient had been previously seen in an IRF, whether an individual was admitted from the community, and FIM cognitive and motor scores at admission. In addition, we controlled for the presence of comorbidities identified by Elixhauser (1998)[2] listed either as part of the ten comorbid conditions in the IRF PAI form or on any acute claim in the prior year. We also controlled for a subset of the complications identified by Iezonni (1994) that may influence post-acute costs and outcomes, either from the preceding acute stay (if one exists) or listed in the ten comorbid conditions on the IRF PAI form, including postoperative pulmonary compromise; postoperative gastrointestinal hemorrhage; cellulitis or decubitus ulcer; septicemia; pneumonia; mechanical complications due to a device, implant, or graft; shock or arrest

[2] Congestive heart failure, valvular disease, pulmonary circulation disorders, peripheral vascular disorders, hypertension, paralysis, other neurological disorders, chronic pulmonary disease, diabetes (uncomplicated), diabetes (complicated), hypothyroidism, renal failure, liver disease, peptic ulcer disease excluding bleeding, AIDS, lymphoma, metastatic cancer, rheumatoid arthritis/ collagen vascular diseases, coagulopathy, obesity, weight loss, fluid and electrolytes disorders, blood loss anemia, deficiency anemias, alcohol abuse, drug abuse, psychoses, and depression.

in the hospital; postoperative myocardial infarction; postoperative cardiac abnormalities other than AMI; venous thrombosis and pulmonary embolism; procedure-related perforation or laceration; acute renal failure; delirium; and miscellaneous complications.

Choice of Models: Random-Effects Versus Fixed-Effects Models

Our first decision was to choose between facility-level random-effects models that assume facility specific effects (quality) are uncorrelated with regressors (i.e., a random-effects model) and facility fixed-effects models that do not require this assumption. Using a Hausman test, we tested the hypothesis that the estimates obtained from fixed-effects and random-effects specifications were equivalent for a model with the base independent variables described above, controlling for initial reason for rehabilitation using RIC. Table 3.1 displays the results of these hypothesis tests; for each of the dependent variables (overall and motor FIM gain, discharge to community, 30-day readmission to acute care, 30-day readmission to SNF, and discharge directly to acute care) we rejected the equivalence of the coefficients in the random- and fixed-effects models, thereby rejecting the assumptions of the random-effects models. Reflecting this, all our main specifications include facility fixed effects, and these estimated facility effects form our quality metric.

Table 3.1

Results of Hausman Test

	Chi-Squared Statistic	P-value
FIM gain	497.76	0.000
FIM gain–motor	851.53	0.000
Discharge to community[a]	561.06	0.000
30-day acute care readmission[a]	204.85	0.000
30-day SNF readmission[a]	511.91	0.000
Discharge directly to acute care[a]	273.63	0.000

[a]To facilitate model convergence, we used a random subsample limited to 200 observations in each IRF and a parsimonious model with simplified comorbidity measures.

Goodness of Fit and Choice of Model Specification

We evaluated an extensive set of models for the final risk adjustment, using all IRF patient stays occurring between 2004 and 2009. The full model includes the full slate of demographic variables, impairment group code for the reason for the rehab stay, comorbidities, complications, admission from community, an indicator for prior IRF stay, and FIM cognitive and motor scores at admission. In addition, we allow for interaction effects between gender and age. Finally, we allowed for interaction effects for having multiple Elixhauser comorbidities in a system and for having Elixhauser comorbidities in more than one system, where the systems are abuse (drug and alcohol), cancer/immune system, cardiac, blood diseases, mental health, neurologic, and weight disorders. We started with this full model and then examined other specifications that omit certain controls, or we switched out controls for more-parsimonious alternatives.

We considered three goodness-of-fit statistics to assess model specifications. For FIM gain, which is continuous, we calculated mean square predicted error for each model, AIC, and BIC. For the other outcomes, which are dichotomous, we calculated the

C-statistic, AIC, and BIC. In each case, we created facility-based estimation and validation samples (two-thirds of patients in each facility were randomly assigned to the estimation sample and one-third was assigned to the validation sample). We estimated the model in the estimation sample, predicted the outcome (e.g. FIM gain) in the validation sample, and then calculated the average squared difference between the prediction and actual in the validation sample (FIM gain) or the C-statistic (discharge to community, 30-day readmission to acute care, 30-day readmission to SNF, and discharge to acute care). Next, we calculated AIC and BIC information criteria from the model estimates using the estimation sample. These statistics take the model log likelihood and penalize for larger numbers of regressors (BIC more than AIC), where small numbers are better.

For each of the models and each of the specifications, we calculated several goodness-of-fit statistics, including the AIC and the BIC. We also calculated the mean squared error for the continuous-outcome model (FIM gain) and C-statistics for the dichotomous outcomes. Each of these statistics measures the extent to which the models explain the variation in the data. The AIC and the BIC include a penalty for inclusion of each additional parameter. Tables A.1–A.6 in the appendix contain these goodness-of-fit statistics for each of the models and each of the specifications. In general, the specifications performed comparably, although the models without initial FIM score performed worse than the other models. Although the ranking of model performance was similar across the goodness-of-fit measures, it was not identical. We therefore chose the best model based on the AIC statistic. For each outcome the best model was either the full model or the model that excluded age-gender interactions.

Determination of Minimum Number of Cases to Support Quality Metric

Estimates of facility quality metrics (i.e., facility fixed effects) become less precise as the number of patients seen in a facility decreases. We assert that the correct metric to determine whether there is enough information regarding a facility's quality is the confidence interval (CI) around the facility's quality metric, which reflects the uncertainty of the quality estimate. We show, however, that this is highly correlated with the number of cases within the facility; therefore, we defined a rule based on the number of cases. To determine the level of uncertainty that is acceptable and the volume of cases

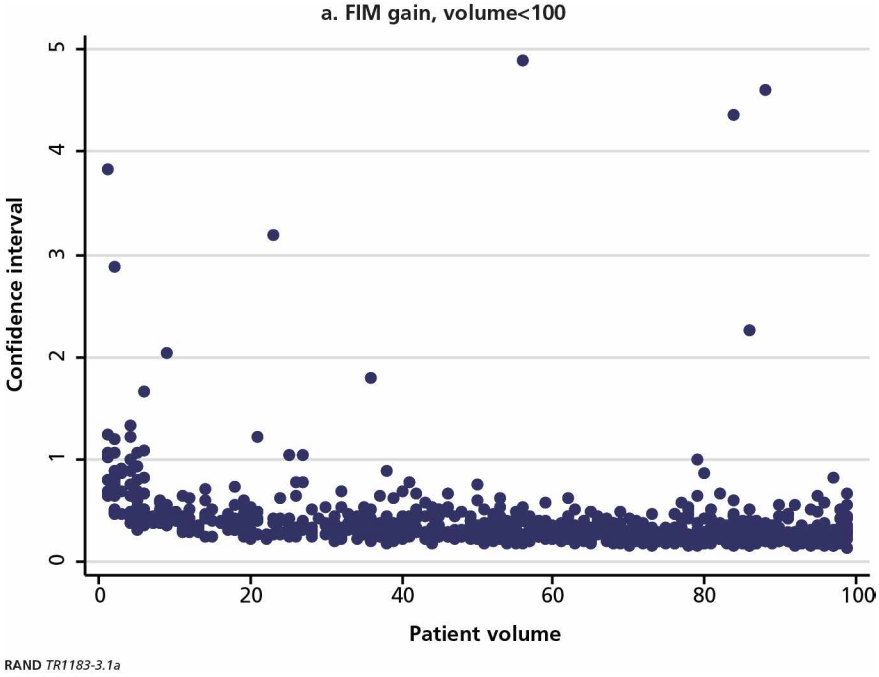
to which this translates, we generated figures of CI plotted against facility patient volume (Figure 3.1). Note that each dot is a CI/volume pair for a particular facility-year, and we limit the figures to those facilities with volume < 100 for FIM gain and < 1,000 for the other dichotomous outcomes. Clearly, there is a strong association between confidence interval and volume, particularly for small patient volumes.

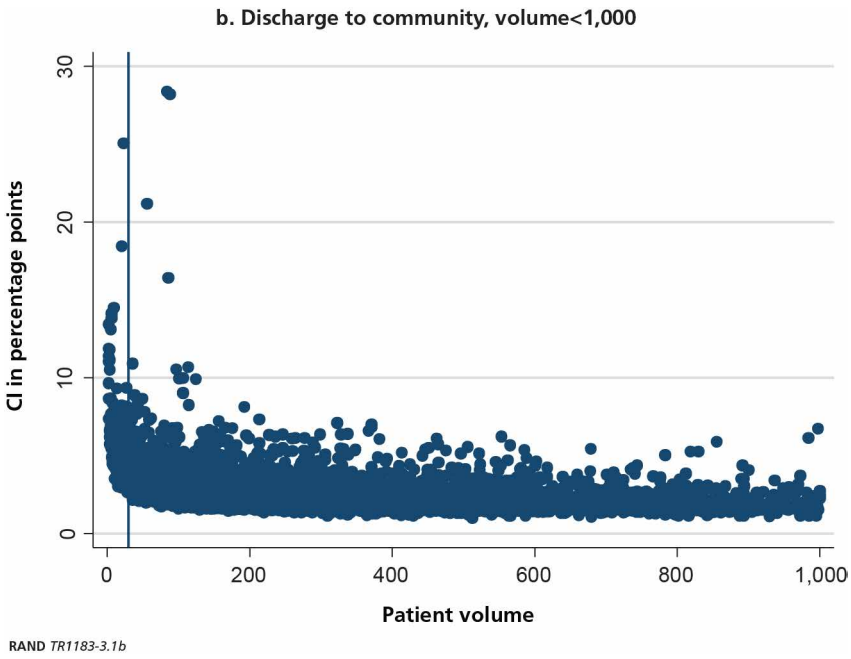
We defined the minimum patient volume to support a quality metric, first by identifying a maximum allowable CI for a facility's quality measure and then by identifying the patient volume at which the CI is always smaller than the maximum allowable CI. We determined the maximum allowable CI using two standards: (1) a suitable CI based on the empirical evidence from Figure 3.1 combined with substantive importance of the variation and (2) the CI determined by one standard deviation in the outcome across IRFs. Because these approaches resulted in very similar thresholds for the CI, and to adopt a uniform approach across outcomes, we adopted the one standard deviation threshold for this study. Because in each model there were several IRFs with relatively large patient volumes and large CIs, we modified the criteria for reporting a Q metric to include both a patient volume threshold (n = 30, to eliminate very small volume IRFs) and a CI threshold (CI of the quality metric > one standard deviation of the outcome across IRFs, to eliminate IRFs with unexpectedly uncertain quality measures).

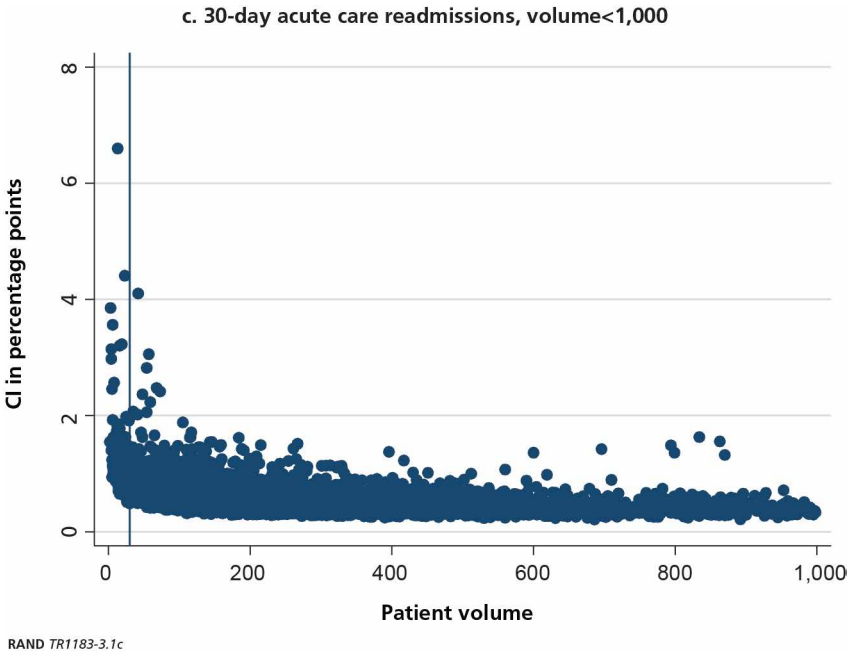
For FIM gain, we considered a confidence interval of 22 (the minimum clinically significant change in overall FIM for stroke patients, as found by Beninato et al. (2006)) and a second maximum confidence interval based on one standard deviation change in FIM as determined in the overall sample (15.67). In the case of FIM gain, however, the confidence intervals for each of the estimated facility fixed effects fell under these thresholds. Thus, for the FIM gain quality measure, we omitted only IRFs with patient volume less than 30.

Figure 3.1

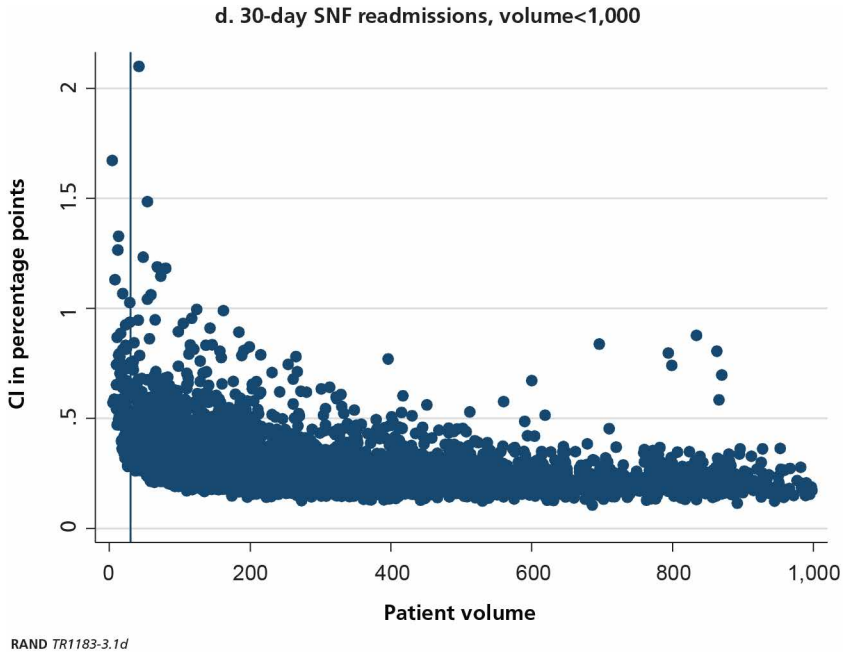
Confidence Intervals by Admission Volume



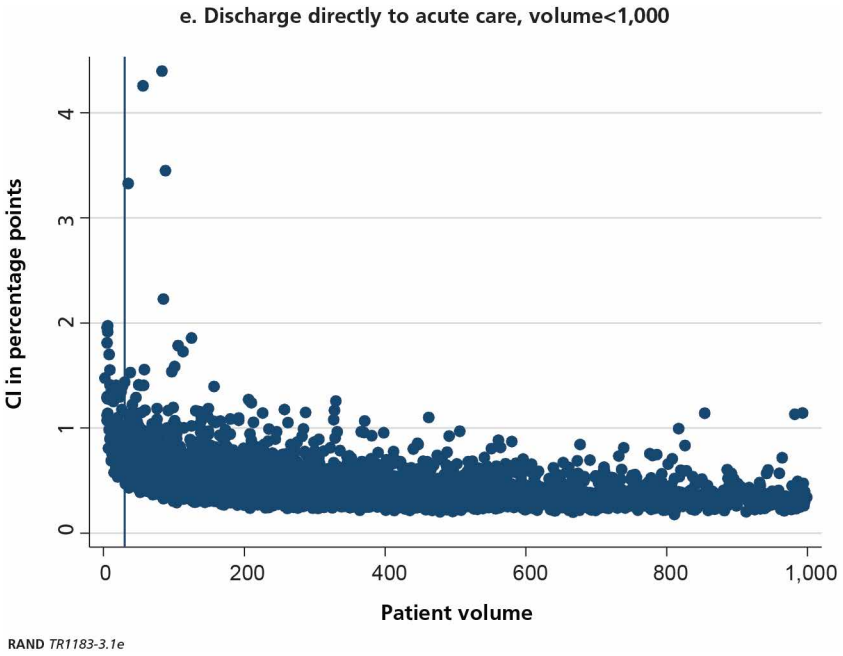




RAND *TR1183-3.1c*



RAND *TR1183-3.1d*



One method for calculating quality metrics for facilities with low patient volume is to combine observations over multiple years of data, thereby increasing the sample size for each institution. Table 3.2 shows the number of facilities (from 2009) that do not meet the inclusion criteria (patient volume < 30 and CI > 1 standard deviation of the outcome) using only 2009 data, and then again, with pooled data over 2 or more years. For example, for FIM gain, the first column shows that 28 (2.4 percent) facilities failed to meet the criteria in 2009. Using multiple years of data, however, fewer of the 2009 facilities were dropped. Only ten, seven and five facilities failed to meet the criteria when we included two years (2008–2009), three years (2007–2009), and four years (2006–2009), respectively. This moving-average approach may require some kind of up-weighting of recent patient stays relative to older patient stays but could lead to suitably precise quality estimates for the vast majority of facilities. The table shows that when pooling three or more years, very few IRFs would be dropped from the reporting sample. To pool over several years, of course, would require assumptions about the stability of the model and each IRF's quality over the time period.

Table 3.2

Omitted Facilities Using a Combined Standard: Volume < 30 or CI > 1 Standard Deviation

	2009 data*	2008–2009 data	2007–2009 data	2006–2009 data
FIM gain	28 (2.4%)	10 (0.86%)	7 (0.60%)	5 (0.43%)
Discharge to community	24 (2.1%)	7 (0.61%)	6 (0.52%)	5 (0.43%)
30-day acute	75 (6.54%)	17 (1.48%)	10 (0.87%)	6 (0.52%)
30-day SNF	49 (4.52%)	14 (1.29%)	6 (0.55%)	3 (0.28%)
Discharge directly to acute	19(1.66%)	6 (0.52%)	4 (0.35%)	3 (0.265%)

*Total numbers of 2009 facilities in the four models are 1,159, 1,154, 1,147, and 1,084, respectively.

Trends in Raw and Adjusted Outcomes

For each of the five outcomes, we estimated year-specific risk-adjustment models and developed year-specific quality metrics. Here we present the raw annual outcome rates and adjusted annual outcome rates for each outcome. To generate the adjusted outcome rates, we used the year-specific risk-adjustment model estimates for each year, predicted the outcomes with each model using the 2004 study cohort, and calculated the mean of the predictions for each year-specific model. This reveals the expected outcome rate for each study year holding the case-mix characteristics constant based on the 2004 study cohort. The raw outcomes rates include the effects of case-mix changes and quality changes. The adjusted outcomes rates characterize the effects of quality changes (holding case mix constant). Thus, a comparison the raw and adjusted rates decomposes the trends into case-mix effects and quality effects.

Table 3.3 presents the results. For each outcome except FIM gain, the raw rates reveal a worsening of outcome rates during the study period. FIM gain improves by about 1.85 points. In each case, however, the adjusted rates show improvement in outcomes during the period. For example, the raw rates of discharge to community declined from 77.75 percent in 2004 to 71.02 percent in 2009. The adjusted rates, however, holding case-mix constant at 2009 sample characteristics, improved from 77.75 percent to 78.92 percent. Thus the increasing severity of the case mix, which drives the trends in the raw rates, masks an overall improvement in quality for each of the indicators.

Table 3.3

Trends in Raw and Adjusted Outcome Rates

	Discharge to Community		30-Day Acute Care Admission		30-Day Long-Term Care Admissions		Discharge Directly to Acute Care		FIM-Gain	
Year	Raw	Adjusted *	Raw	Adjusted *	Raw	Adjusted *	Raw	Adjusted *	Raw	Adjusted *
2004	77.75	77.75	10.83	10.83	3.11	3.11	8.67	8.67	25.25	25.25
2005	75.44	77.94	10.47	9.83	3.19	2.98	9.08	8.14	25.77	25.88
2006	73.18	78.13	11.02	9.72	3.34	2.91	9.39	7.61	25.95	26.29
2007	72.06	78.51	11.36	9.61	3.52	2.88	9.92	7.61	26.19	26.79
2008	71.31	78.38	11.38	9.39	3.63	2.90	10.14	7.55	26.55	27.25
2009	71.02	78.92	11.58	9.31	3.72	2.90	10.25	7.22	27.10	27.86

* Adjusted outcome rates adjust population characteristics to match the 2004 sample.

FIM Gain

Stability of Coefficient Estimates over Time

As described above, we tested the stability of the model parameter estimates by comparing 2004 estimates to those from a pooled sample from 2006–9, and by comparing 2006–7 estimates to those from 2008–9. In both cases, we strongly rejected the equivalence of the model estimates ($P<0.01$), suggesting that the models changed throughout the study period and that caution should be used when pooling several years of data to develop IRF quality metrics. There are several possible explanations for the changing parameter estimates, including (1) unobserved heterogeneity in severity not picked up by the covariates, masking real differences in the study population over time, (2) changes in technology (e.g., equipment, knowledge, medications) such that the risks associated with the covariates change over time, or (3) changes in IRF practice patterns.

Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates

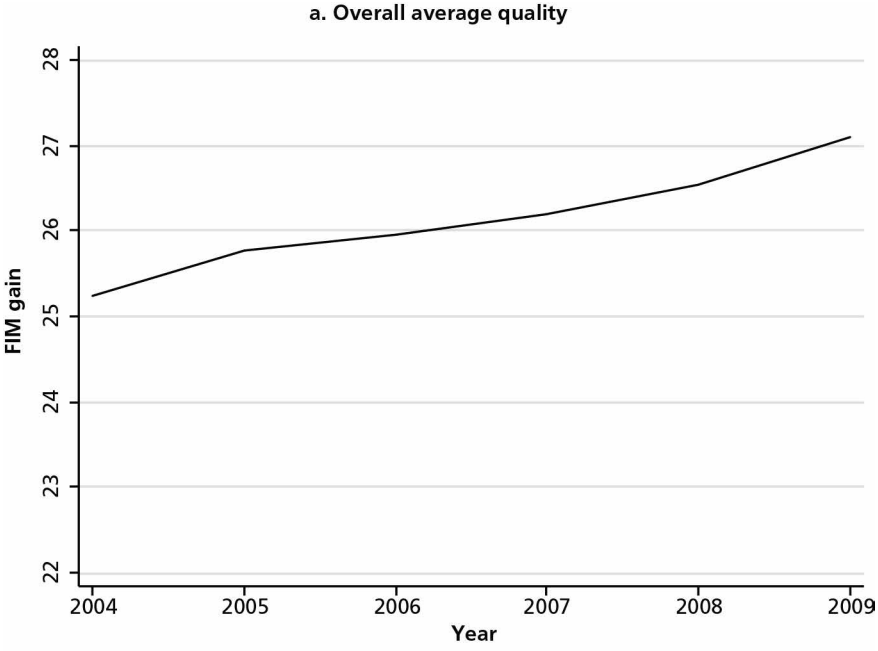
Figure 3.2a shows that the average FIM gain increased from around 25 to 27 over the sample period. This increase could be caused by two factors: changes in the composition of patients (case mix) and changes in risk-adjustment parameter estimates—possibly representing changes in practice patterns. Below, we estimate separate risk-adjustment models to investigate the impacts of the two factors.

In Figure 3.2b, we use risk-adjustment parameter estimates for each year to predict FIM gain for a fixed cohort of patients. Thus, year- to-year changes in average FIM are restricted to changes from the risk-adjustment parameter estimates and not the composition of patients. Each line represents predicted FIM gain for a different cohort of patients. The upward slope of each line implies that quality, as reflected by FIM gain as a function of patients' observable characteristics, has improved over the sample period. That is, over time the outcomes are better for identical patients. The monotonic decrease over time in the intercept of each patient cohort's line implies that the patient case mix has worsened over time. These results may be the consequence of the implementation of the 75 percent rule. After implementation, patients' health status may have declined, which would result in a decreasing intercept, indicating that they had a reduced potential for rehabilitation.

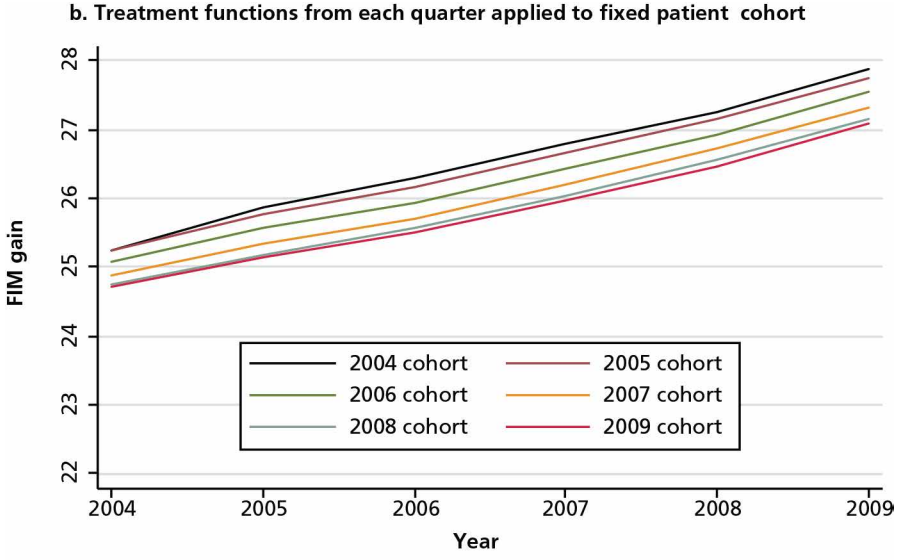
In Figure 3.2c, we use risk-adjustment parameters from a single year to predict FIM gain for each annual cohort of patients—in this case, year-to-year changes in average FIM are restricted to changes from the composition of patient case mix, not the regression estimates. The downward slope of each line implies a negative patient composition effect on outcomes: Patients in later cohorts have characteristics associated with less FIM gain. These composition effects are consistent with the effects of the 75 percent rule. IRFs were required to admit patients with more-severe disabilities who may represent less potential for functional improvement relative to simple joint replacement patients. However, holding patient composition constant, quality, as measured by FIM gain in Figure 3.2b, has actually improved.

Figure 3.2

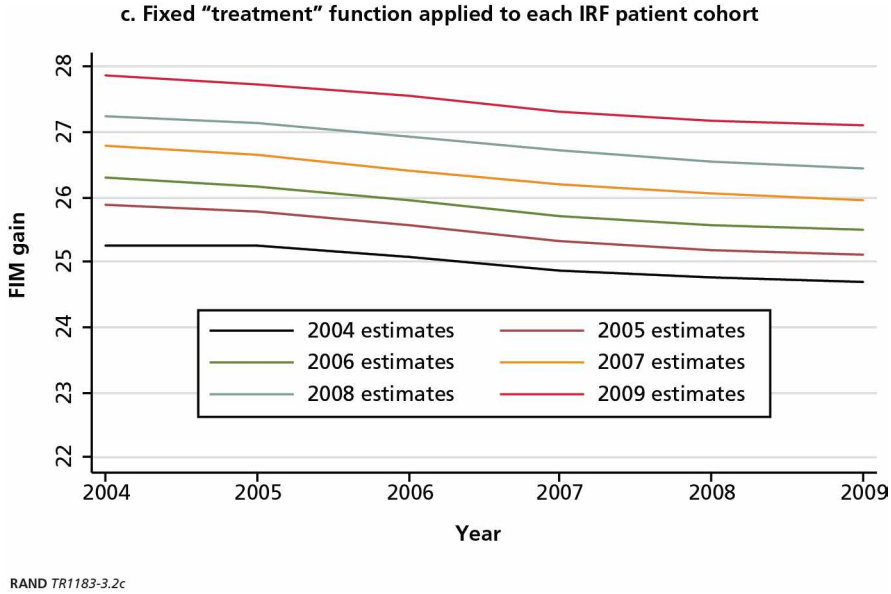
Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates (FIM gain)



RAND *TR1183-3.2a*



RAND *TR1183-3.2b*



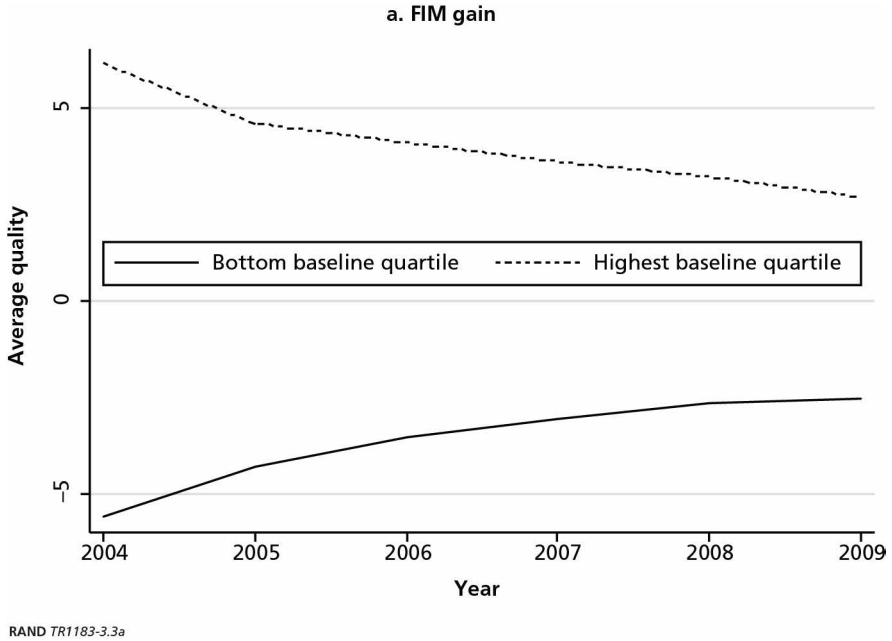
Measures of Stability of Quality

Next, we examine trends in IRF quality metrics separately by provider characteristics. We estimated separate risk-adjustment models (using the full model outlined in the description of Table 3.1) to create a panel of IRFs with quality metrics for each year, where the quality metric is the facility fixed effect in each year. We then linked these data with provider characteristics from point-of-service (POS) data to examine differential levels and trends in quality across facilities.

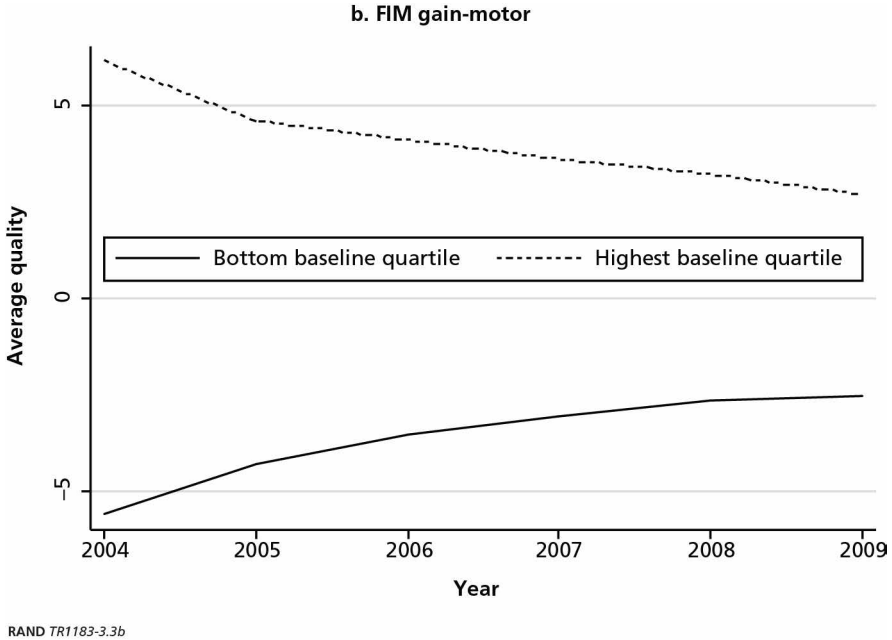
Figure 3.3 shows trends in quality for IRFs in the lowest and highest quartile of quality in 2004. There is some convergence in the highest and lowest quartiles, suggesting some degree of mean reversion. However, the highest quartile stays above mean quality and the lowest quartile stays below mean quality on average, suggesting that our quality metrics are capturing persistent differences in quality.

Figure 3.3

High Versus Low Baseline Performance, 2004 IRF Cohort (FIM gain)



RAND *TR1183-3.3a*



RAND *TR1183-3.3b*

Discharge to Community

Stability of Coefficient Estimates

As described above, we compared the model parameter estimates using 2004 data to those of 2006–9, and we compared estimates from the 2006–7 sample to those of 2008–9. In both cases, we strongly rejected the equivalence of the parameter estimates, indicating that the models changed over time and that caution should be taken when using pooled data to estimate risk-adjustment models. In both of the comparisons (2004 versus 2006–9 and 2006–7 versus 2008–9), many of the parameter estimates were similar, but some were substantially different (e.g., age, gender, marital status, race/ethnicity, dual eligibility for Medicare and Medicaid, and admission from the community), and the test of joint equivalence was strongly rejected ($P<0.01$).

Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates

Figure 3.4a shows that the rate of discharge to the community decreased over time, from 78 percent to 71 percent in 2009. However, this does not necessarily mean that quality of care decreased, since patients may have become more severely impaired in recent years.

When we used a fixed cohort and applied each year's model estimates, Figure 3.4b shows a similar pattern to that of FIM gain. We found that the predicted rate of discharge to the community actually increased over time, but the magnitude was relatively small, less than two percentage points. This indicates that quality of care as measured by the rate of discharge to community increased from 2004 to 2009. The monotonic decrease in the intercept of each cohort shows that the case mix worsened over time, resulting in a reduction in discharge to the community.

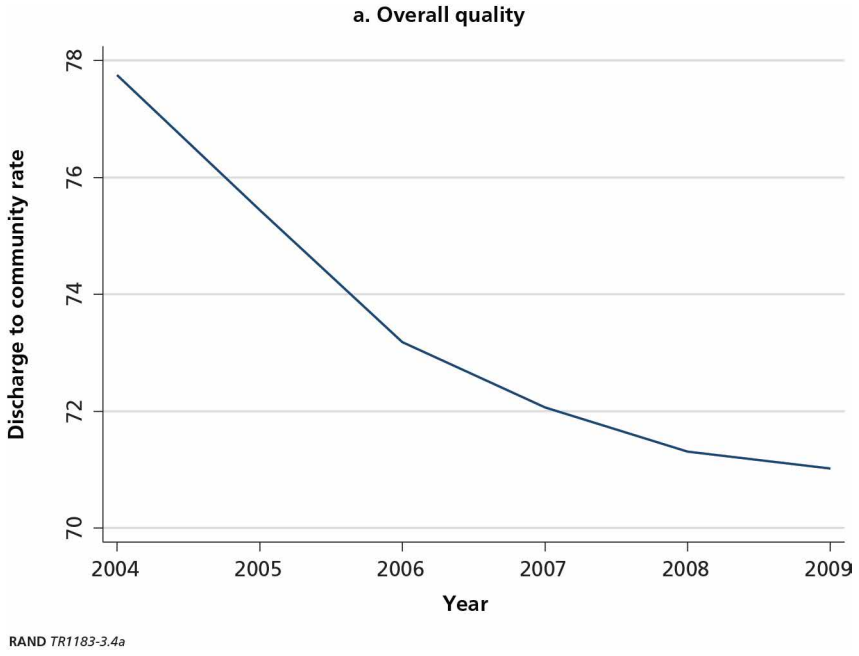
Figure 3.4c further confirms our findings from Figure 3.4b. By fixing the model estimates, the trend lines represent the impact of patient case-mix characteristics, with the more recent patient cohorts more severely impaired than those in earlier years. In addition, the magnitude of the decline in the rate of discharge to the community is similar to that of Figure 3.4a, implying that a majority of the change in the overall outcome may

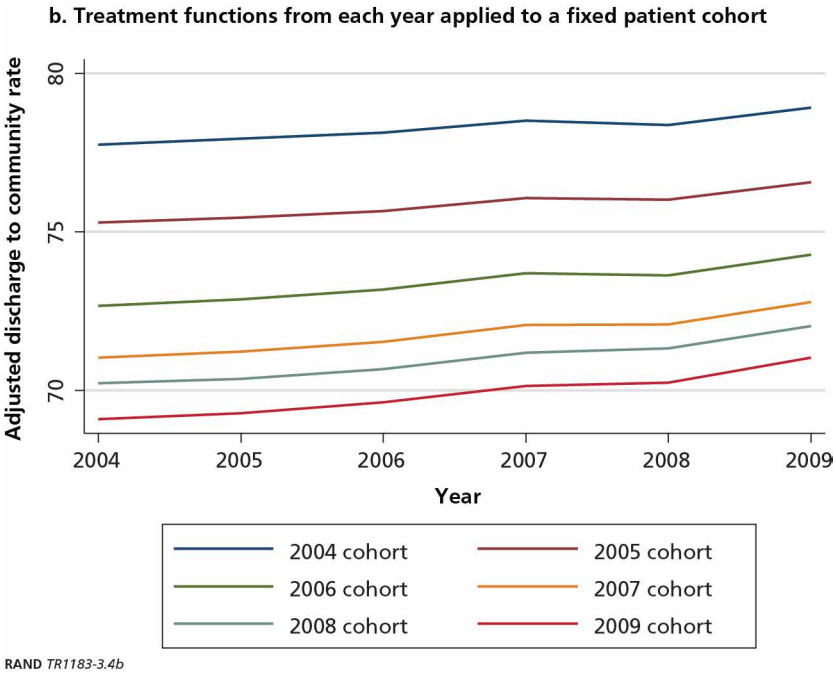
be due to patient composition rather than quality of care. The difference in intercepts of Figure 3.4c represents the improvement in quality, but the magnitude is only about two percentage points.

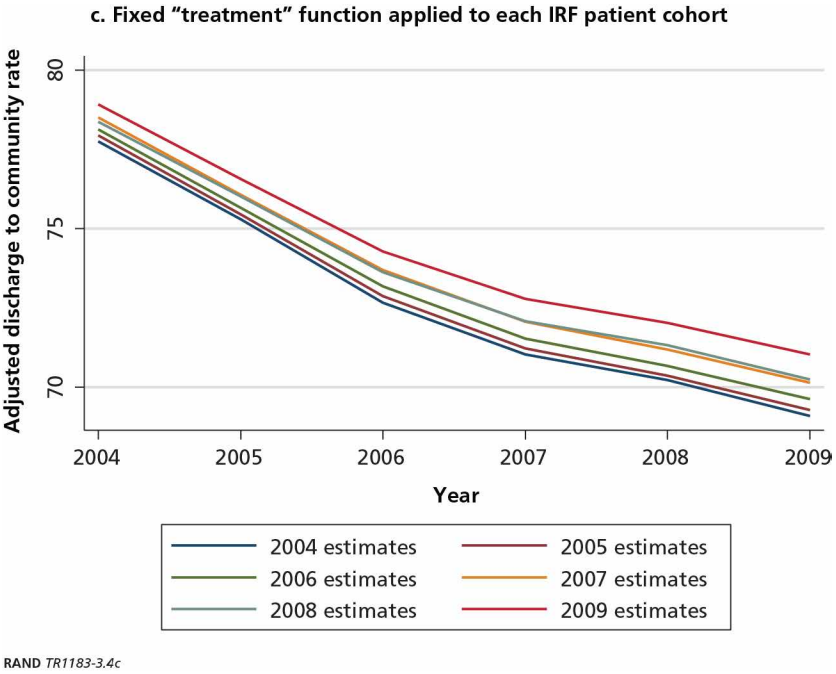
The quality of care, as measured by the rate of discharge to community, has been improving over time, but the impact of patient composition (worsening case mix) overwhelms the effect of quality improvement. Thus we observe a declining trend in the overall rate of discharge to community as shown in Figure 3.4.

Figure 3.4

Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates (Discharge to Community)







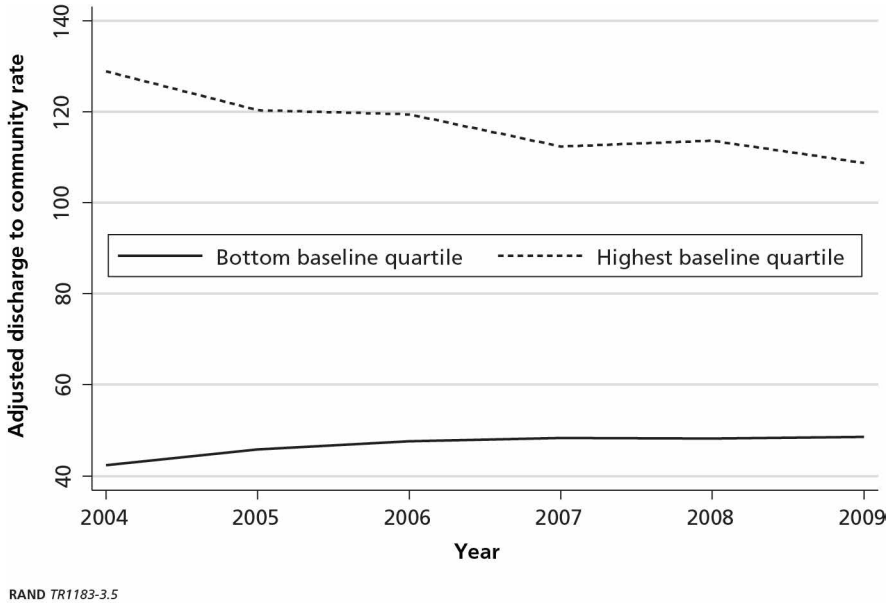
Measures of Stability of Quality

Figure 3.5 shows the trend in risk-adjusted rates of two IRF quality cohorts over time. The IRFs in the highest quartile of performance in 2004 experienced a decline in quality while the opposite is true for those in the lowest quartile of performance in 2004. The

high-quality cohort consistently outperforms the low-quality cohort over time, although the difference in quality diminishes.

Figure 3.5

High Versus Low Baseline Performance, 2004 IRF Cohort (30-Day SNF Readmissions)



30-Day Acute Care Readmission

Stability of Coefficient Estimates

As described above, we compared model estimates using a 2004 sample to those made with a 2006–9 sample. We also compared model estimates using a 2006–7 sample to those using a 2008–9 sample. Note that because of difficulties achieving model convergence, we estimated a more parsimonious specification for the second comparison, substituting RIC for IGC and simplifying the structure of the comorbidity controls. Not surprisingly, the 2004 estimates appeared to be most dissimilar from the later years, perhaps as a consequence of the rule change. We found important differences by age, race, dual eligibility status, prior IRF experiences, and IGCs. There were few differences by comorbid conditions, including the number and type of systems involved. The joint test of the equivalence of 2004 model estimates and 2006–9 model estimates resulted in a chi2 (with 153 degrees of freedom) statistic of 350.84 ($P<0.001$). Differences in the later

years (2006–7 versus 2008–9) were similar but smaller. The joint test of equivalence over the two periods resulted in a chi2 (with 77 degrees of freedom) equal to 127.59 ($P<0.001$).

Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates

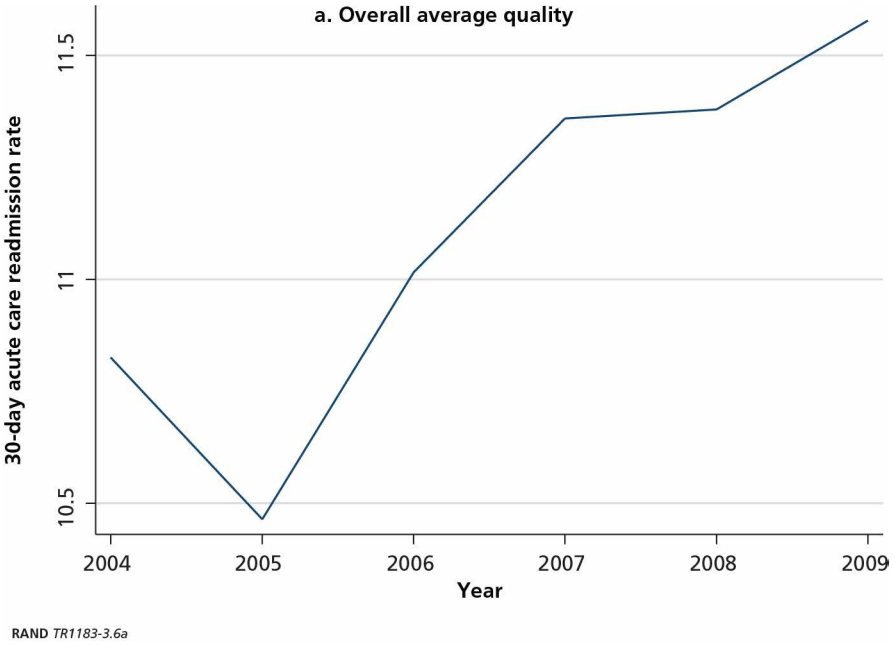
For 30-day acute care readmissions, we found that the observed rate dropped from 2004 to 2005 and then had an upward trend for the rest of the study period, as shown in Figure 3.6a. This implies that the overall quality among IRFs declined during the period, but it does not control for case mix.

We also used each year's data to estimate year-specific risk-adjustment parameters, and we applied the estimates to predict 30-day acute care readmissions for fixed cohorts of patients, defined by year in our sample. Figure 3.6b shows a downward slope for each line, suggesting an improvement in the 30-day acute care readmission over the study period if the composition of patients is held constant. Because the curve for each successive year is higher, the figure also shows that the population mix has become higher-risk over time.

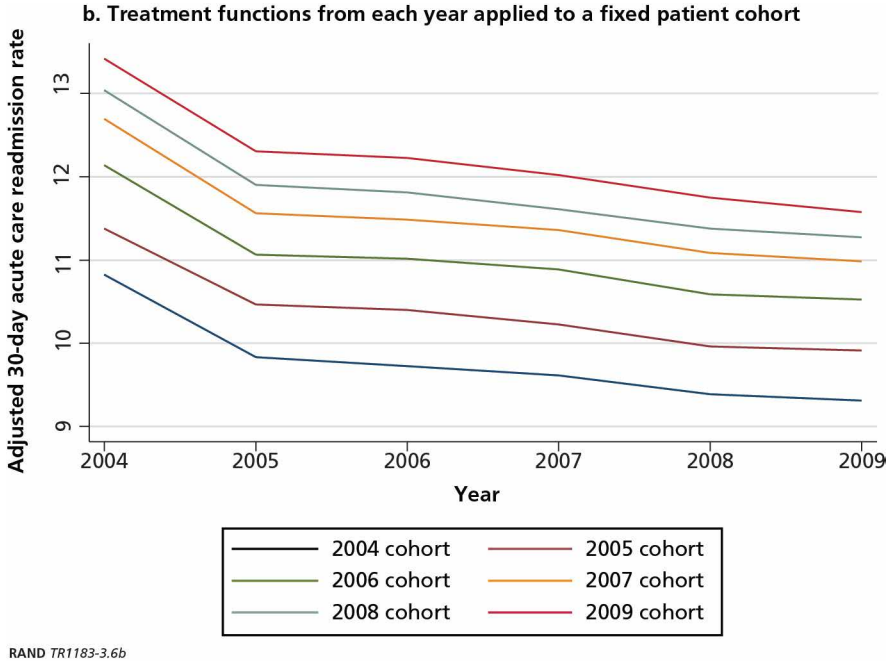
In Figure 3.6c, we use risk-adjustment parameters from a single year to predict the 30-day acute care readmission rates for each annual cohort of patients. The upward slope of each line indicates a positive patient composition effect on readmission rates, implying that patients in later cohorts have characteristics associated with higher 30-day acute care readmission rates.

Figure 3.6

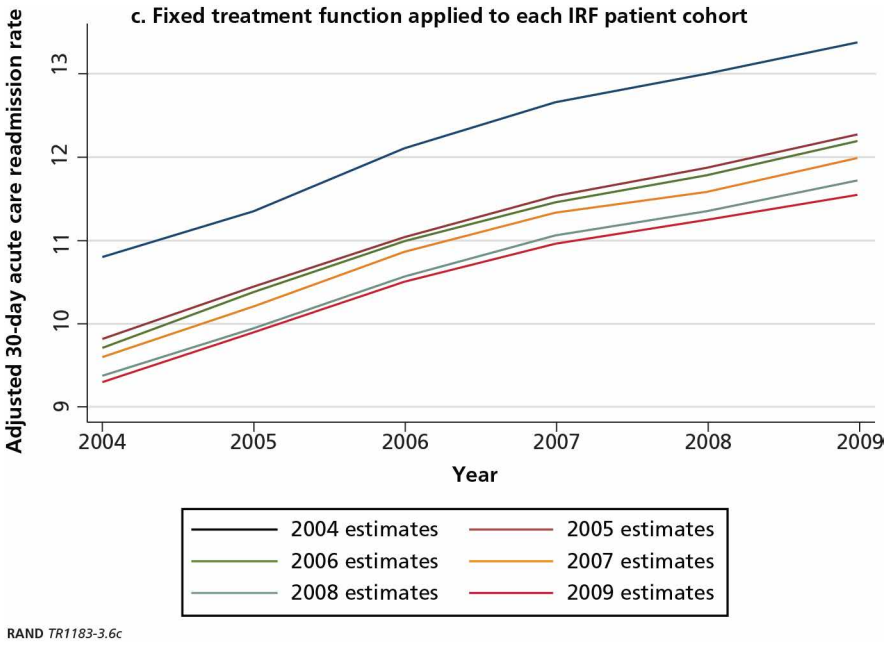
Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates (30-Day Acute Admissions)



RAND *TR1183-3.6a*



RAND *TR1183-3.6b*

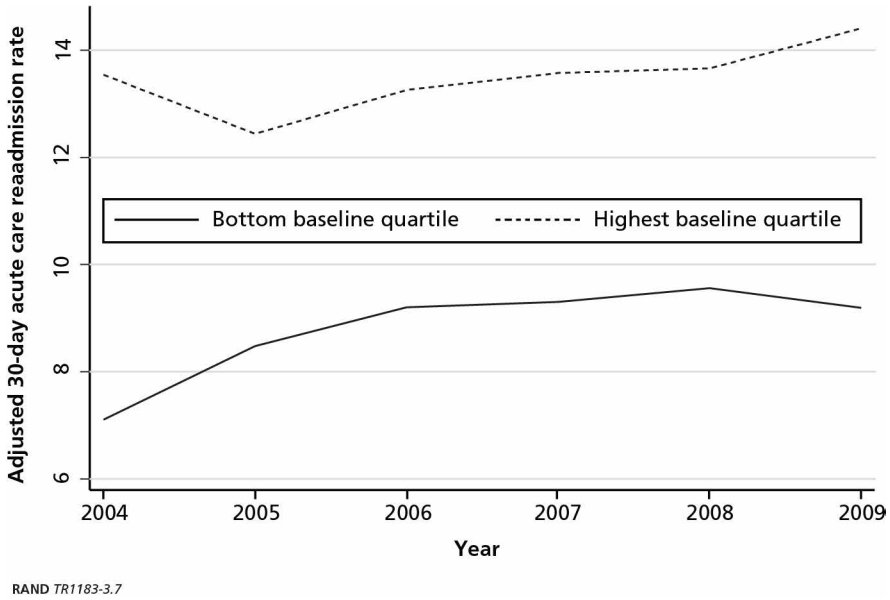


Measures of Stability of Quality

Figure 3.7 shows the trend in the risk-adjusted rate of two IRF quality cohorts over the study period. Note that the cohorts are defined by the 2004 quality status and not the status in subsequent years. Therefore, the figure shows the persistence of the 2004 quality difference over time. The rate for the IRFs in the highest quartile (lowest quality) of outcomes in 2004 first declined (improved) in 2005, and then increased (worsened) for the rest of the study period. In comparison, the rate of those IRFs in the lowest quartile (highest quality) of outcomes in 2004 had an upward trend, except in 2009, the last year of the study period. Overall, the high-quality cohort (lowest quartile) consistently fared better than the low-quality cohort (highest quartile) over time. By the end of the study period, the difference between the two cohorts had become even larger.

Figure 3.7

High Versus Low Baseline Performance, 2004 IRF Cohort (Discharge Directly to Acute Care)



30-Day SNF Readmission

Stability of Coefficient Estimates

As with the outcomes described above, we made two comparisons over time (2004 versus 2006–9 and 2006–7 versus 2008–9) to examine the stability of model estimates. As with the 30-day acute care readmissions, we estimated a more parsimonious specification for the second of these comparisons in order to facilitate model convergence. We used RIC indicators rather than IGC indicators, and we simplified the comorbidity measures. We also used a subsample in which we randomly drew a maximum of 400 observations for each IRF in each period. Unlike the other outcomes, however, the results for the 30-day SNF readmissions outcome were not consistent. We found large and statistically significant differences between estimates from 2004 and 2006–9. As above, these differences were notable in age, dual eligibility status, and selected IGCs. There was little evidence of differences in the comorbidity estimates. The

test of the joint equivalence of the parameter estimates resulted in a chi2 (148) equal to 226.95.

The comparison of model estimates in the two later periods (2006–7 versus 2008–9), however, was quite different. To facilitate model convergence, we simplified the models by replacing the IGC indicators with RIC indicators and by simplifying the structure of the comorbidity measures. Of the 77 parameters that we estimated separately by time period, we found statistically significant differences (using a critical value of 0.05) only for marital status and a single RIC. The test statistic for the joint equivalence of the parameter estimates was chi2 (77) equal to 68.56 (P= 0.74). Thus, at least during the last four years of the study sample, the model coefficient estimates were stable.

Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates

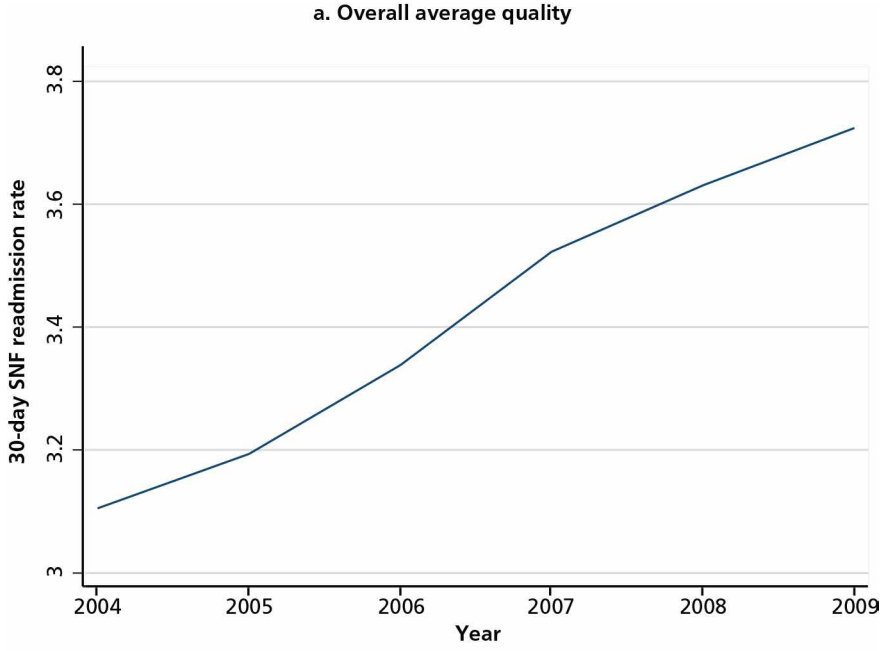
The overall 30-day SNF readmission rate increased monotonically from about 3.1 percent to 3.7 percent during the study period (Figure 3.8a). As with the 30-day acute care readmissions outcome, this figure shows that the outcome markedly worsened on average during the study period, but it does not control for case mix.

The figures that control for case mix show a markedly different story. Figure 3.8b, which shows the average IRF quality for fixed cohorts of subjects (defined by the year of our sample), shows that readmissions fell monotonically during the study period, implying an improvement in quality. Each successive year's cohort, however, had higher readmission rates throughout the study period, indicating a worsening of the case mix. The worsening case mix overwhelmed the improvement in quality, resulting in the overall increase in the outcome rate during the study period.

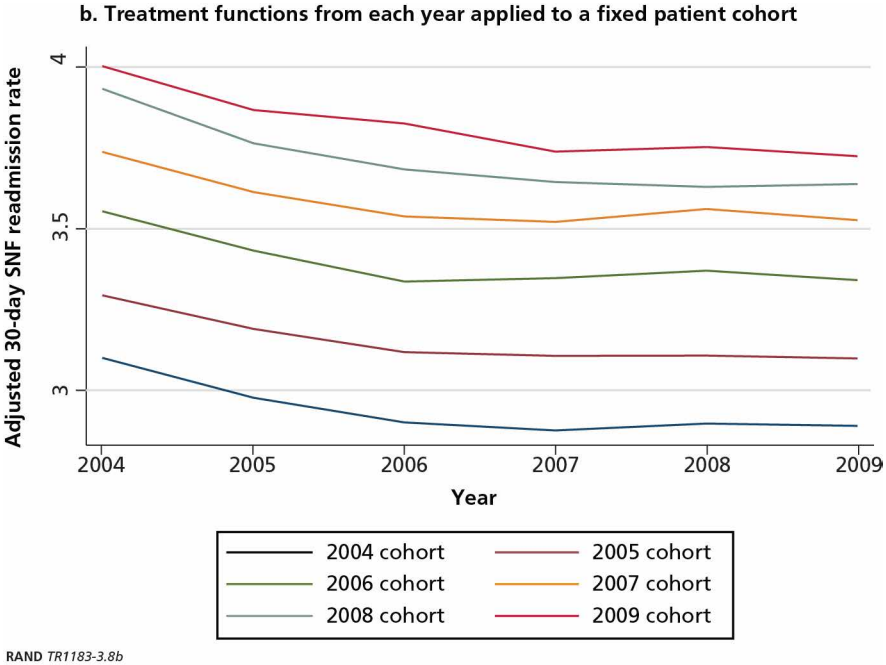
In Figure 3.8c, we used risk-adjustment parameters from a single year to predict the 30-day SNF readmission rates for each annual cohort of patients. The upward slope of each line indicates a positive patient composition effect on readmission rates, implying patients in later cohorts have characteristics associated with higher 30-day SNF readmission rates.

Figure 3.8

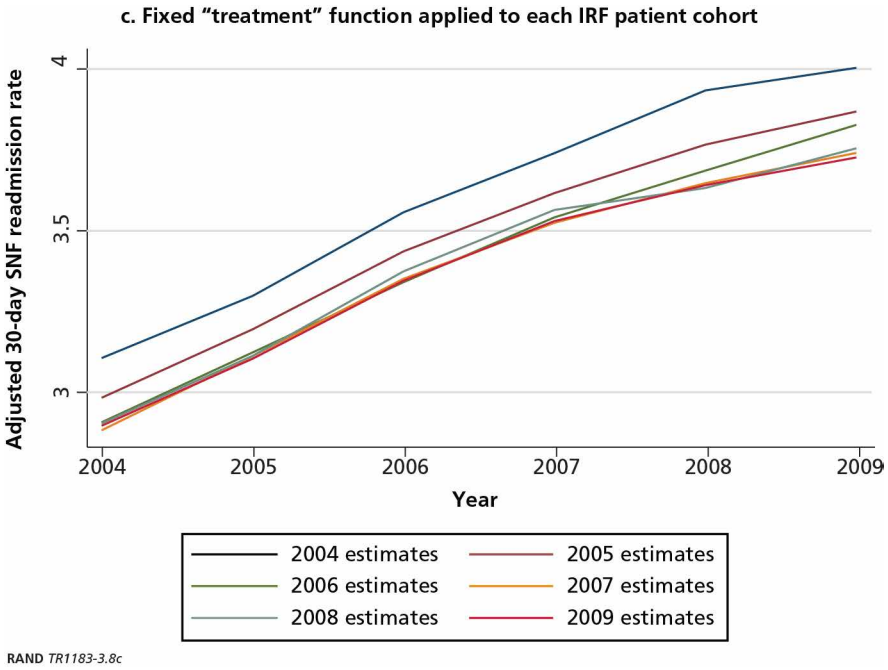
Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates (30-Day SNF Readmissions)



RAND *TR1183-3.8a*



RAND *TR1183-3.8b*

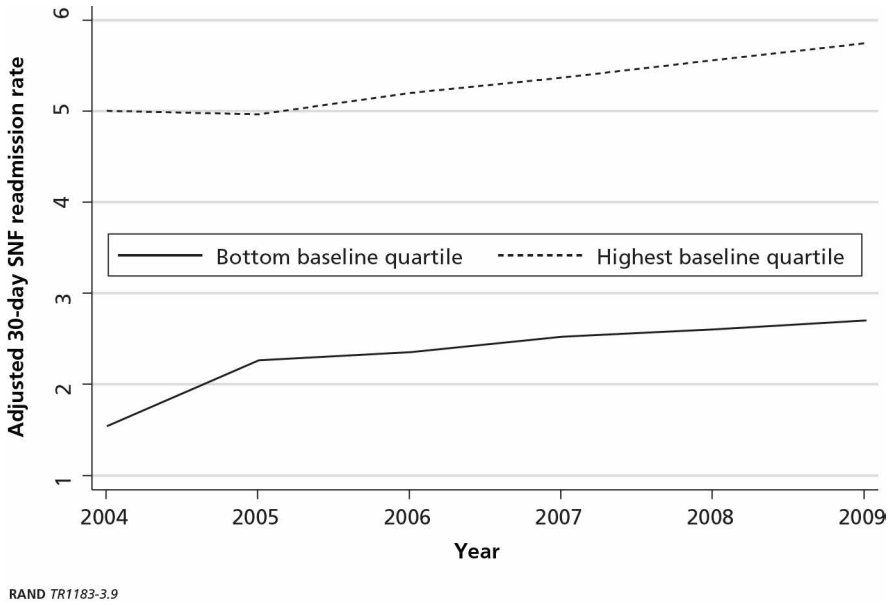


Measures of Stability of Quality

Figure 3.9 shows the trend in the risk-adjusted rate of two IRF quality cohorts over the study period. As described above, the cohorts are defined by the 2004 quality status and not the status in subsequent years. Therefore, the figure shows the persistence of the 2004 quality difference over time. The rate for the IRFs in the highest quartile (lowest quality) of outcomes in 2004 held steady in 2005, and then increased at a steady rate. In comparison, the rate of those IRFs in the lowest quartile (highest quality) of outcome in 2004 had an abrupt increase in 2005 and then held relatively steady during the remainder of the study period. The notable feature of the figure, however, is not the overall increase in the outcome (as described above, this is almost entirely due to a worsening case mix), but rather the fact that the higher performers and the low performers (as identified in 2004) maintained their difference in performance throughout the study period. We did not observe a regression toward the mean. This could imply that IRF quality based on SNF readmission rates is persistent.

Figure 3.9

High Versus Low Baseline Performance, 2004 IRF Cohort (30-Day SNF Readmissions)



Discharge Directly to Acute Care

Stability of Coefficient Estimates

As above, we compared model estimates for 2004 versus 2006–9 and for 2006–7 versus 2008–9. In both cases, to facilitate model convergence we estimated parsimonious specifications in which we replaced the IGC indicators with RIC indicators and simplified the structure of the comorbidity estimates. We found evidence that the model coefficient estimates differed over time. In particular, we found differences in 2004 versus 2006–9 in dual eligibility status, prior IRF experiences, RICs, and the number of comorbidities (by body systems). We soundly rejected the joint equivalence of the parameter estimates (chi2 (77) = 156.50; P<.001). We found differences in the 2006–7 versus 2008–9 parameter estimates for marital status, race, selected RICs, and the number of comorbidities (by body systems). The joint test of the equivalence of the parameter estimates strongly rejected equivalence (chi2 (77) = 139.35; P<0.001).

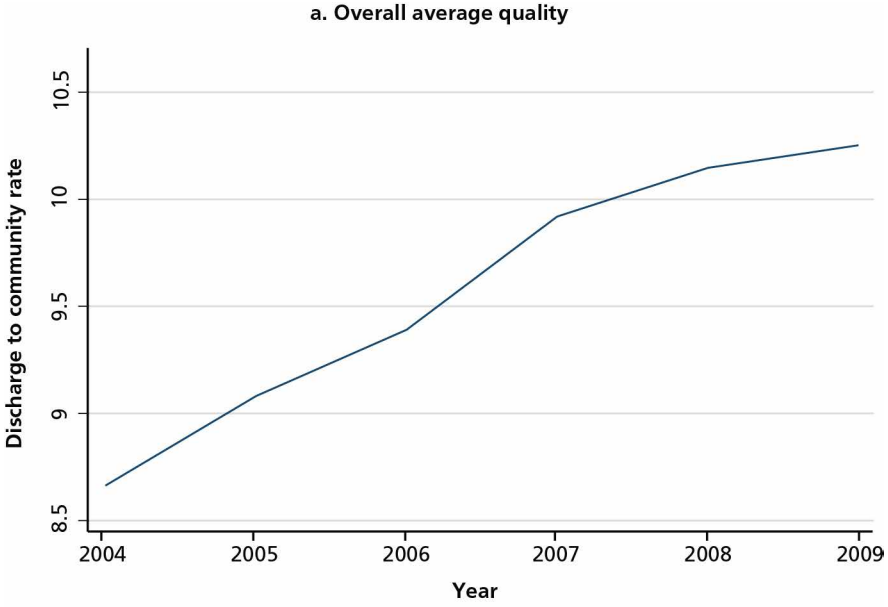
Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates

The overall outcome rate for discharge directly to acute care increased steeply from just over 8.5 percent to about 10.25 percent during the study period (Figure 3.10a). The overall picture painted by the raw outcome rate is one of steadily worsening industry-wide quality, but this does not control for case mix, and—as we found with the other outcomes—this fact is important.

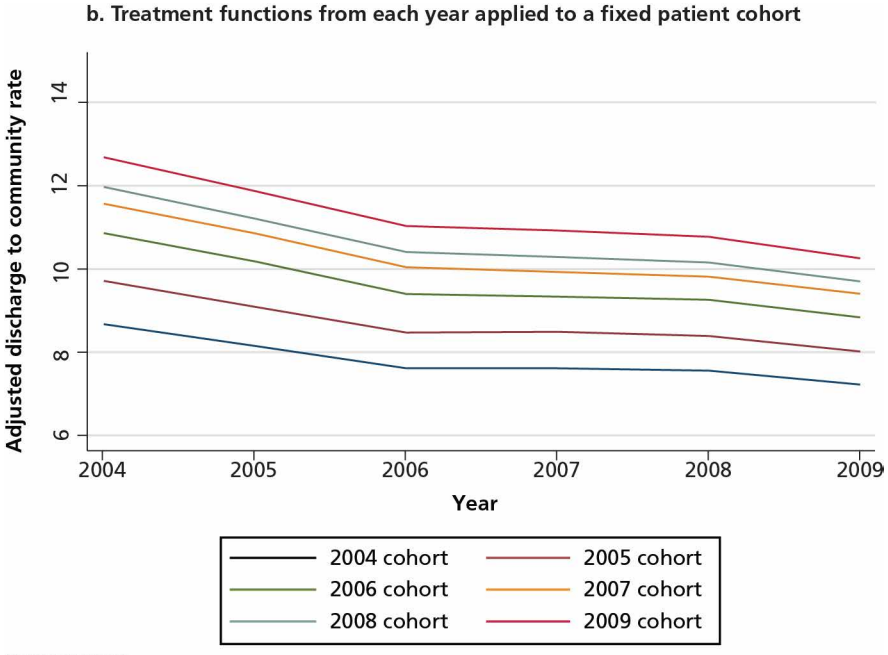
Figure 3.10b shows the average IRF quality (based on discharge directly to acute care) for fixed cohorts of subjects defined by the year of our sample. Each line, therefore, represents a different (year-specific) cohort and shows how the outcome rate changes over time for a fixed cohort of patients. The figure shows that acute care admissions fell monotonically during the study period after controlling for case mix. Each successive cohort (different lines by year), however, had higher predicted readmission rates throughout the study period. This indicates a steadily worsening case mix that explains the overall increasing trend in the outcome indicator. As with the readmission indicators, the worsening case mix obscures an overall improvement in quality in the industry.
In Figure 3.10c, we used risk-adjustment parameters from a single year to predict the rates of discharge directly to acute care for each annual cohort of patients. The upward slope of each line indicates a positive patient composition effect on readmission rates, implying patients in later cohorts have characteristics (case mix) associated with higher outcome rates (discharge to acute care).

Figure 3.10

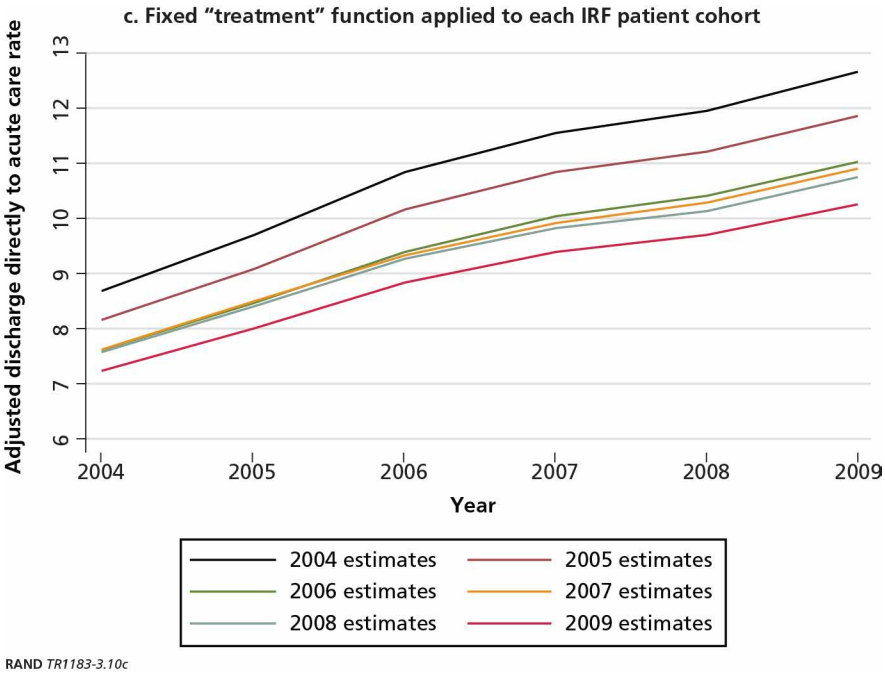
Trends in Quality and Decomposition of Trends into Population Composition Versus Changes in Risk-Adjustment Estimates (Discharge Directly to Acute Care)



RAND *TR1183-3.10a*



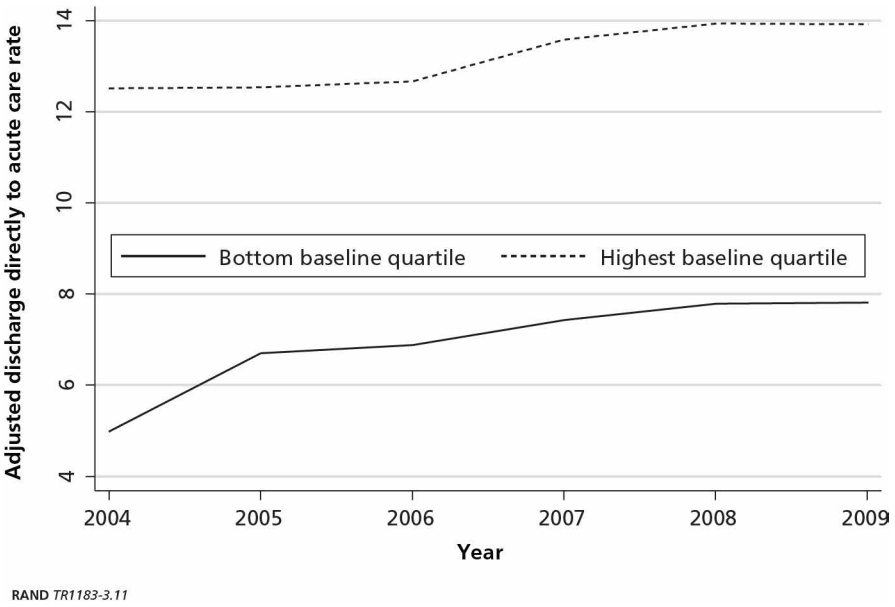
RAND *TR1183-3.10b*



Measures of Stability of Quality

Figure 3.11 shows the trend in the risk-adjusted rate of two IRF quality cohorts over the study period. As described above, the cohorts are defined by the 2004 quality status and not the status in subsequent years. Therefore, the figure shows the persistence of the 2004 quality difference over time. The rate for the IRFs in the highest quartile (lowest quality) of outcomes increased only slightly during the study period. In comparison, the rate of those IRFs in the lowest quartile (highest quality) of outcome also increased slightly during the period, resulting in a persistent gap between the quartiles.

Figure 3.11
High Versus Low Baseline Performance, 2004 IRF Cohort (Discharge Directly to Acute Care)



4. DISCUSSION

We used a consistent approach to develop, estimate, and test risk-adjustment models for five patient outcomes: (1) FIM gain, (2) discharge to community, (3) 30-day readmission to acute care, (4) 30-day readmission to a SNF, and (5) discharge directly to acute care. For each outcome, we began by specifying a model that contained individual socioeconomic and demographic characteristics, comorbid condition indicators, IGC indicators, and age-by-sex interactions. We estimated nine alternative specifications for each model and chose the model with the best AIC score. In each case, the best model was either the full model or the model that dropped the age-by-sex interactions. Model fit was good, but there remained a substantial amount of unexplained outcome variation by IRF.

We considered two main approaches for risk-adjusting to model IRF quality: (1) measures based on random effects or hierarchical models and (2) measures based on fixed-effects models in which each IRF's contribution to quality is determined by a fixed indicator in the model. For each of the outcomes models, we found strong evidence to reject the random-effects assumption, indicating that quality measures based on random effects estimates would be biased. We therefore adopted the fixed-effects approach. In the case of this application, even though the number of IRFs was large (1,400), the very large data set (2.5 million observations) made it feasible to estimate fixed-effects models. One notable advantage of the fixed-effects approach is that the models generate a nonparametric empirical representation of the distribution of quality across IRFs. In addition, by manipulating the models so that each IRF's estimate represents its difference from the average quality IRF, the quality metrics are relatively easy to manipulate and interpret.

Although our study sample contained approximately 2.5 million observations, not every IRF had large patient volume. Thus, we developed criteria for determining if the quality estimate for an IRF was too uncertain or unreliable to merit reporting. Starting with the assumption that the estimated uncertainty (standard error or confidence interval

of the quality measure) is the best means for determining the measure's reliability, we examined the relationship between IRF volume and the quality measure's uncertainty, and we examined various criteria that used only IRF volume to determine the omission of IRFs. We found that the use of a volume standard either resulted in the omission of too many IRFs (many of which had reliable estimates of quality) or the inclusion of IRFs that had very uncertain quality estimates. We therefore recommend the use of a combined standard that requires a minimum patient volume (we used n = 30) to eliminate very small volume IRFs *and* a maximum uncertainty level (confidence interval > one standard deviation of the outcome rate across institutions) to exclude IRFs for which the quality estimates were not precise. We found that this combination standard resulted in the exclusion of a very small number of IRFs (see Table 3.2 above).

We examined model stability by comparing model estimates based on data from early in the study period to model estimates based on data from late in the study period. In every case but one (30-day readmission to SNF comparing 2006–7 estimates to 2008–9 estimates), we found evidence that the model parameter estimates changed over time. This implies that caution should be used when making model estimates using data pooled over several years. It also suggests that risk-adjustment models will need to be recalibrated regularly. As a result, we estimated year-specific risk-adjustment models to generate year-specific quality indicators for each IRF. These indicators can be compared either within year to reveal any IRFs quality relative to other IRs or the average IRF or over time to reveal trends in quality.

There are several important limitations of this study. First, no risk-adjustment models are perfect. Although we used state-of-the art risk-adjustment tools here, there could be important unobserved heterogeneity in case severity that could generate biased parameter estimates and compromise the IRF quality estimates. This issue is particularly important for this study given the revealed strong trend toward worsening case-mix during our study period. In addition, the ability of the IGC and RIC controls to account for the 2004 regulatory change is particularly important for the face validity of the study. Changes in unobserved severity, for example, could be responsible for the differences in the case-mix parameter estimates that we uncovered.

Second, we found that, in addition to a substantial increase in case-mix severity during the study period, there were important differences in the case-mix parameter estimates during the period. Although we addressed this by estimating year-specific models, this adds a layer of complexity to the interpretation of quality trends over time. The IRF fixed effects parameter estimates from a given year (and metrics derived from them) reveal IRF relative quality only for that year. Comparisons across years provide information only on how a given IRFs relative quality (in a given year) changed from year to year. To uncover real trends in quality, we developed an additional metric that accounted for the changes in the case-mix parameter estimates as well as the relative quality (within year) of each IRF. This standardized prediction metric, therefore, facilitates comparisons over time, and, although it is computationally burdensome to calculate, we used this approach. In addition, care must be taken when making comparisons over time because the interpretation and validity of such comparisons will depend on assumptions. These include, for example, whether the differences in the risk-adjustment models over time is caused by changes in technology, and thus real changes in quality, or changes in unobserved case-mix.

Third, although this study used a rich and very large data set, there were individual IRFs that had few observations in a given year. We found that inclusion criteria based on a combination of a volume threshold and an estimation uncertainty standard resulted in the exclusion of a small subset of IRFs. Quality estimates for many of these IRFs could be made by pooling data over two or more years, but this would limit the ability to identify year-specific quality and, therefore, changes in quality from one year to the next. In addition, although we rejected the stability of the case-mix parameter estimates, it would be cumbersome (and perhaps infeasible) to specify year-specific case mix parameters because of the computational burden of estimating these models.

Fourth, our study made no effort to address the quality of rehabilitation care provided in other settings. Given the substantial increase in case-mix severity observed in our study, it is likely that care provided in other settings has become increasingly important. Finally, although we considered five outcome measures (FIM gain, discharge to the community, discharge directly to acute care, 30-day long-term care admissions, and 30-

day acute-care admissions), these may not adequately cover the dimensions of IRF quality. In addition, we did not examine the extent to which quality metrics generated from these five outcomes are positively or negatively correlated and whether it would be feasible to develop a single quality index.

We used the models to examine trends in quality during the study period. Among the key findings:

1. *The raw outcome rates for four of the five quality measures worsened over the study period*. FIM gain improved by about two points, but discharge to the community (18 percent to 11 percent) 30-day readmission to acute care (10.5 percent to 11.5 percent), 30-day readmission to SNF (3.1 percent to 3.7 percent), and discharge directly to acute care (8.6 percent to 10.25 percent) all worsened. (See Figures 3.2a, 3.4a, 3.6a, 3.8a, and 3.10a.)
2. *The declining raw rates are caused by worsening case mix during the period*. After adjusting for case mix, we found that quality improved on every metric and, in the case of FIM gain, quality improved by more than the improvement in the raw rate. (See Figures 3.2b, 3.4b, 3.6b, 3.8b, and 3.10b.)
3. *IRF quality persisted over the study period*. We identified two cohorts of IRFs using 2004 quality estimates: the lowest-quartile performers and the highest-quartile performers. We then examined the average quality of those cohorts in each of the study years, and we found that the 2004 high-performing IRFs continued to be high performers throughout, that the 2004 low performers continued to be low performers, and that the gap in quality did not narrow substantially. (See Figures 3.3, 3.5, 3.7, 3.9, and 3.11.)

Our work indicates that the existing data provide ample information to estimate risk-adjustment models for the purpose of examining and reporting IRF quality. Furthermore, our findings show that, because of the substantial worsening of case mix, the use of high-quality risk-adjustment models as the basis for quality reporting is essential for understanding the relative quality across institutions and for revealing overall quality trends in the market.

Appendix

MODEL SPECIFICATIONS

Table A.1

Model Specifications, FIM Gain

Model Description	Mean Square Error (validation sample)	AIC	BIC
1. Full model	227.64	13833265	13835536
2. Full model, no complications	227.98	13835759	13837882
3. Full model, ten year-age indicators	227.83	13835117	13837277
4. RIC indicators instead of IGC	228.36	13838106	13839587
5. No race, married, dual eligibility	227.88	13834336	13836545
6. No age-gender interactions	227.64	13833262	13835422
7. Omit dummy variables for numbers of diseases in each system	227.65	13833366	13835488
8. Omit dummy variables for numbers of systems	227.66	13833309	13835444
9. Omit initial FIM cognitive and motor	233.60	13855516	13857775
10. No system variables and no age-gender interactions	227.67	13833439	13835303

Table A.2

Model Specifications, FIM Gain–Motor

Model Description	Mean Square Error (validation sample)	AIC	BIC
1. Full model	116.32	12697694	12699965
2. Full model, no complications	116.50	12700281	12702404
3. Full model, ten year-age indicators	116.42	12699608	12701755
4. RIC indicators instead of IGC	116.72	12703153	12704647
5. No race, married, dual eligibility	116.43	12698650	12700860
6. No age-gender interactions	116.32	12697691	12699851
7. Omit dummy variables for numbers of diseases in each system	116.33	12697799	12699922
8. Omit dummy variables for numbers of systems	116.33	12697737	12699872
9. Omit initial FIM cognitive and motor	118.63	12722605	12724863
10. No system variables and no age-gender interactions	116.34	12697883	12699746

Table A.3

Model Specifications, Discharge to Community

Model Description	C-Statistic	AIC	BIC
1. Full model	0.7828	1,597,104	1,599,387
2. Full model, no complications	0.7820	1,599,414	1,601,537
3. Full model, ten year-age indicators	0.7615	1,598,709	1,600,733
4. RIC indicators instead of IGC	0.7618	1,598,649	1,600,796
5. No race, married, dual eligibility	0.7780	1,610,945	1,619,893
6. No age-gender interactions	0.7827	1,597,199	1,599,371
7. Omit dummy variables for numbers of diseases in each system	0.7827	1,597,125	1,599,260
8. Omit dummy variables for numbers of systems	0.7827	1,597,326	1,599,473
9. Omit initial FIM cognitive and motor score	0.7052	1,766,545	1,768,866

Table A.4

Model Specifications, 30-Day Acute Care Readmissions

Model Description	C-Statistic	AIC	BIC
1. Full model	0.6795	815269.8	817580.7
2. Full model, no complications	0.6773	816252.4	818394.8
3. Full model, ten year-age indicators	0.6789	815590.5	817022.8
4. RIC indicators instead of IGC	0.6790	815570.6	817135.3
5. No race, married, dual eligibility	0.6788	815640.2	817866.8
6. No age-gender interactions	0.6795	815261.4	817451.9
7. Omit dummy variables for numbers of diseases in each system	0.6794	815314.5	817456.9
8. Omit dummy variables for numbers of systems	0.6796	815347.1	817489.5
9. Omit initial FIM cognitive and motor score	0.6779	816031.6	818318.4

Table A.5

Model Specifications, 30-Day SNF Readmissions

Model Description	C-Statistic	AIC	BIC
1. Full model	0.6985	343248.0	345462.4
2. Full model, no complications	0.6976	343371.5	345477.6
3. Full model, ten year-age indicators	0.6982	343348.0	344756.1
4. RIC indicators instead of IGC	0.6978	343406.0	344946.4
5. No race, married, dual eligibility	0.6943	344214.0	346428.3
6. No age-gender interactions	0.6984	343245.2	345339.2
7. Omit dummy variables for numbers of diseases in each system	0.6984	343308.0	345799.2
8. Omit dummy variables for numbers of systems	0.6984	343288.9	345370.9
9. Omit initial FIM cognitive and motor score	0.6848	345811.0	348073.5

Table A.6

Model Specifications, Discharge Directly to Acute Care

Model Description	C-Statistic	AIC	BIC
1. Full model	0.7167	943789.9	946047.2
2. Full model, no complications	0.7137	946380.7	948477.6
3. Full model, ten year-age indicators	0.7047	944856.3	946200.8
4. RIC indicators instead of IGC	0.7047	944863.3	946331.2
5. No race, married, dual eligibility	0.7167	943819.9	946003.2
6. No age-gender interactions	0.7167	943806.0	945952.4
7. Omit dummy variables for numbers of diseases in each system	0.7166	943899.9	946009.3
8. Omit dummy variables for numbers of systems	0.7168	943911.0	946045
9. Omit initial FIM cognitive and motor score	0.6638	983405.6	985749.3

REFERENCES

Beninato, M., K. M. Gill-Body, S. Salles, et al. (2006). "Determination of Clinically Important Difference in the FIM Instrument in Patients with Stroke." *Archives of Physical Medicine and Rehabilitation* 87(1): 32–39.

Carter, G. M., O. Hayden, S. M. Paddock, and B. O. Wynn (2003). *Case Mix Certification Rule for Inpatient Rehabilitation Facilities*. Santa Monica, Calif.: RAND Corporation, DRU-2981-CMS. As of June 5, 2012: http://www.rand.org/pubs/drafts/DRU2981.html

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey (1998). "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36(1): 8–27.

Iezzoni, L. I., J. Daley, T. Heeren, et al. (1994) "Identifying Complications of Care Using Administrative Data." *Medical Care* 32(7): 700–715.

MedPAC (2010). "Rehabilitation Facilities (Inpatient) Payment System. Payment Basics." Washington, D.C. As of June 5, 2012: http://www.medpac.gov/documents/medpac_payment_basics_10_irf.pdf

MedPAC (2011). *Report to the Congress: Medicare Payment Policy*. Chapter 9: Inpatient Rehabilitation Facility Services. Washington, D.C. As of June 5, 2012: http://www.medpac.gov/documents/Mar11_EntireReport.pdf