Identification and Evaluation of Existing Quality Indicators that are Appropriate for Use in Long-Term Care Settings

CMS Contract No:
500-95-0062/T.O. #4

Final Report

October 1, 2001

Prepared for

Yael Harris, Project Officer
Centers for Medicare and
Medicaid Services
Office of Clinical Standards and
Quality / S3-02-01
7500 Security Boulevard
Baltimore, Maryland 21244

Principal Authors

Katherine Berg
Brant E. Fries
Richard Jones
Soeren Mattke
Terry Moore
Vincent Mor
John N. Morris
Katharine M. Murphy
Sue Nonemaker

Contract Agencies:
Abt Associates Inc. (*Prime Contractor)*
55 Wheeler Street
Cambridge, Massachusetts 02138-1168

HRCA Research and Training Institute
1200 Center Street
Roslindale, Massachusetts 02131

Brown University
Center for Gerontology and Health Care Research
171 Meeting Street, Box G-B213
Providence, Rhode Island 02912

University of Michigan
Institute of Gerontology
300 N. Ingalls
Ann Arbor, Michigan 48109

Key Staff:
John N. Morris
Principal Investigator, HRCA
Katharine M. Murphy
Co-Investigator, HRCA
Vincent Mor
Co-Investigator, Brown University
Katherine Berg
Co-Investigator, Brown University
Terry Moore
Project Manager, Abt Associates

CMS Project Officers:
Sue Nonemaker/Yael Harris

Lead Agency for this Project Activity:
Brown University

Internal Review

Project Director

Technical Reviewer

Management Reviewer

Contents

Technical Appendices:

Appendices:

Acknowledgments

The Project Team is grateful to the contributions and insights of many others who tested and reviewed existing quality indicators in the following domains:

Clinical -
Katherine Berg
Chris Brostrup-Jensen
Caroli Flodstrom
David Gifford
Rich Jones
Alex Kartashov
Lew Lipsitz
Courtney Lyder
Ed Marcantonio
John Morris
Katharine Murphy
Sue Nonemaker
Joan Teno
Tom Wachtel

Drug -
Chris Brostrup-Jensen
Giovanni Gambassi
David Gifford
Sharon Gonzalez
Ann Hume
Alex Kartashov
Kate Lapane

Drug (cont'd) -
Lew Lipsitz
Vince Mor
Sue Nonemaker

Functional -
Joseph Angelelli
Katherine Berg
Ann Deutsch
Brant Fries
Carl Granger
Dan Kiely
John Morris
Shirley Morris
Katharine Murphy

Psychosocial -
Barbara Bell
Chris Brostrup-Jensen
Dan Kiely
Susan Miller
Terry Moore
Vincent Mor
John Morris

The team would also like to acknowledge the following project consultants and technical expert panel members who commented on the methods and findings of this report:

Consultants -
Sara Karon
David Zimmerman

Technical Expert Panel Members -
Dana Mukamel
Marilyn Rantz
Barbara McNeil

Technical Expert Panel (cont'd) -
John Hirdes
Pam Duncan

Reviewers -
David Zimmerman
Christina Bethel

Executive Summary

The Center for Medicare and Medicaid Services (CMS), like many other health insurance payors, has begun to demand more accountability on the part of its Medicare service providers such as hospitals, nursing facilities, home care agencies, managed care organizations (MCOs) and physicians. One aspect of accountability of particular interest to CMS is that of patient care outcomes. Despite some measurement and interpretation difficulties, care outcomes may now be assessed remotely, via electronic database surveillance systems, for select providers of service. This makes affordable and timely information on care quality available to CMS as both a purchaser of service and regulator of quality care for Medicare beneficiaries. For example, Medicare MCOs are monitored via the National Committee on Quality Assurance measures. Nursing facilities have recently begun to be monitored via quality indicators (QIs) constructed from the resident assessment tool used by all United States nursing facilities to generate the Minimum Data Set (MDS). Medicare-certified home care agencies will soon be monitored for their care outcomes via "Outcome-based QIs", derived from the newly-mandated OASIS assessment instrument.

As CMS moves further toward monitoring the performance of providers through performance measurement systems, there is a need to develop and test additional quality measures. In developing new QIs and revising existing measures, it will be important that care quality be assessed for the full range of Medicare beneficiaries and other long-term care users. For example, unique subsets of the population that access long-term care services may have very different care needs and care outcomes from other subsets; thus, these special populations must be considered as new QIs are constructed. CMS awarded this contract, entitled *The Development and Validation of Long-term and Post-acute Care QIs*, in order to obtain new quality measures for the long-term (or chronic) and post-acute populations that use nursing facility services. In addition, the project will assess whether quality of care provided to special populations, such as the palliative care population, can be measured with standard QIs or whether their special needs require the development of targeted QIs.

Prior to developing new QIs, the first task of the project was to assess existing QIs and to determine which of them, if any, can be recommended to CMS for immediate use. While previous work has focused primarily on how QIs could augment the regulatory process, the focus of this project is to identify QIs for use by multiple audiences, ranging from nursing facilities themselves to the consumers of and purchasers of care. It is now expected that nursing facilities will increasingly use information regarding their performance to improve quality, and that consumers may start incorporating information about quality into their decisions. Similarly, purchasers may, in the future, base contracting and other decisions on facility performance as measured by QIs. All four listed audiences — facilities, regulators, consumers and purchasers — have been considered in the work described in this report. This report summarizes work conducted for this task by describing the process of searching for QIs in the literature and the results derived from evaluating identified QIs. It concludes with explicit recommendations on the use of existing QIs.

Identification of Existing QIs

An extensive review of published and unpublished literature on QIs used in the long-term and post-acute care populations was conducted. Researchers in the field and QI system vendors were contacted regarding the existence of unpublished or proprietary QIs. This extensive search, described in detail in Chapter 3, yielded a total of 143 QIs. Clinical and other staff from the cognizant project agencies (Abt Associates, Brown University Center on Gerontology, Hebrew Rehabilitation Center for the Aged, and the University of Michigan) reviewed all QIs and categorized them as primarily functional, clinical, psychosocial or pharmacotherapy QIs. In addition, QIs were classified as either prevalence-based or longitudinal, and by the data source used to construct them (e.g., MDS, medical record data).

Following abstraction, all 143 QIs were evaluated against a set of criteria. Criteria included having an explicit operational definition (i.e., a defined numerator and denominator), being constructed from MDS or similar data, and having some form of risk adjustment. If a QI was found to meet these criteria, it was passed forward to be empirically reviewed. In addition, all the QIs that the Center for Health Systems Research and Analysis (CHSRA) had developed for the CMS survey and certification system were moved forward to the empirical analysis. A total of 44 QIs underwent further empirical analyses. The process is outlined in Chapter 4.

Empirical Analyses Conducted on Existing QIs

Initially, the 44 selected QIs were examined using a longitudinal file of MDS data from five states, to determine:

- the distribution of raw and adjusted QI rates within and across all five states;
- the relationship between raw and adjusted QI rates; and
- the consistency of the QI rates across states and time periods.

Additional analyses were conducted in order to understand the inter-relationships among QIs and their potential vulnerability to problems such as differential censoring, casemix differences, and misclassification by facility assessors. The methods for those analyses are described in Chapter 5 and the results in Chapters 6.

Based upon those initial analyses, the project team identified a total of 26 QIs as being sufficiently reliable and valid measures of care quality for nursing facilities in their internal quality improvement efforts. However, without additional refinement, none of them were judged suitable for public reporting to consumers or purchasers, and only a few of them were deemed to be appropriate for guiding the survey process in an unbiased manner. The main concern was that those existing QIs were not adequately adjusted for differences in casemix and in assessment accuracy across facilities, rendering them vulnerable to selection and ascertainment bias. Empirical analyses supporting this concern are described in the Technical Appendix 1.

Therefore, as described in Chapter 7, the project team recommended modifications to the covariate structures originally applied by the various QI developers for all of these 26 QIs. The modifications entail reducing the list of covariates for some of the QIs (e.g. those designed by the vender LTCQ Inc.). Most importantly, the project team recommended that all QIs be adjusted by introducing a facility-level variable that accounts for differences in admission practices as well as differences in facilities' assessment practices as seen in the profile of residents admitted to the facilities.

The 26 QIs were empirically reevaluated after application of the refined method of risk adjustment, using MDS V2.0 data from Massachusetts. This process resulted in a final recommendation of 22 QIs for use by CMS. Several of these 22 QIs are already in use nationally, such as the "prevalence of daily physical restraint use" QI developed by CHSRA. Others (e.g., those developed by LTCQ, Inc.) would be new QIs to CMS's regulatory oversight program, if adopted.

Recommendations on the Use of QIs for Different Audiences

As noted, the project was undertaken specifically to explore the feasibility of applying existing quality indicators to the various uses different audiences might make of them. Enlightened facilities throughout the country are already using readily available MDS data to target clinical problems where professional staff believe that improvement in patients' status may be achieved. Using QIs for internal quality improvement and monitoring purposes avoids the worst of the technical problems associated with comparing facilities because there's unlikely to be a major change in the type of residents entering the facility or in how resident assessments are done. Thus, comparisons of quarter to quarter changes within a facility can shed light on how well quality interventions might be improving care and outcomes. Regulators are being instructed to rely upon QIs to guide the implementation of the facility survey by helping them to focus on identifiable quality problems. Since QI reports rank homes viz. each QI, there is an implicit inter-facility comparison made which is subject to differences in the mix of residents being admitted as well as the homes' approach to resident assessment. Rankings based upon inadequate adjustment could influence regulators' approach to a facility during the survey. Nonetheless, regulators do have the opportunity to amend their impression of the home based upon their experience surveying the facility.

In contrast to facility CQI efforts or even regulatory uses of QIs, consumers or purchasers will use this information to "screen" facilities for selection without ever seeing those facilities. Optional QIs for use by these audiences should, therefore, be more rigorous in their design. They should be able to be considered "absolute" markers of quality, rather than measures which require further information (e.g., on-site inspection) to fully understand.

After having empirically examined the performance of each of the 22 QIs, the project team assessed how well the 22 finally recommended QIs would serve these four different audiences. The conclusion was that all 22 QIs, even without additional modifications of their risk adjustment procedure, could be used by nursing facilities for purposes of internal quality improvement. Only with the newly-adopted risk adjustment process could the 22 QIs be recommended for use in the regulatory process and 15 of them can be considered for communicating facility performance to consumers or purchasers. As described above, this more cautious stance on public reporting of QIs is warranted.

Summary of Additional Analyses and Recommendations for Further Research

A main concern in the implementation of an indicator-based quality reporting system is that judgments based on those QIs might be influenced by facility characteristics other than quality of care. The project team investigated the impact of casemix differences resulting from differential admission or discharge practices and of differential ascertainment as the most likely sources for such biased assessments. The results showed that this concern is warranted and that the specification of appropriate risk adjustment models is a key requirement for the validity of any QI. Other analyses conducted reveal that, particularly in smaller facilities, rankings based on QIs may vary substantially over time and, therefore, that statements about QI performance cannot be made with much statistical confidence.

The findings presented in this report suggested two main avenues to remedy these problems. The first is the development of methods that adjust for differences in resident risk across facilities. The second involves the incorporation of facility characteristics into the construction of QIs. An initial step in this direction was made by the development of a new risk adjustment method. Ultimately, as a potential approach to these challenges, the project team is studying the use of hierarchical modeling techniques and presents some preliminary results here. This technique appears suitable for the problem at hand, as it allows one to account for resident-level and facility-level characteristics simultaneously in a unified statistical model. For the time being, the less precise approach of treating facility-level adjusters as if they were measured at the resident level tends to generate results that are quite comparable to the hierarchical model.

1.0 Overview and Objectives for Review and Validation of Existing QIs

The present project is designed to assist CMS in advancing their vision for stimulating quality of care in nursing facilities by developing and validating QIs that reflect clinical and other care outcomes at the facility level. Underlying this vision is the characterization of a facility's performance by examining resident assessment information during a specific time frame. The vision also assumes that providing feedback on performance and holding facilities accountable for their performance will lead to improvements in care outcomes. That is, facilities will strive to perform as well as or better than other facilities in the marketplace and will avoid receiving demerits or deficiency scores from surveyors. Public presentation of comparative rankings of performance or deficiency scores will create incentives for improvements in the quality of care.

Comparison of nursing facility performance based on actual resident or patient outcomes requires the presence of specific building blocks. For example, facilities must use the same measures in assessing and in monitoring outcomes. Since 1991, federal regulations have mandated that nursing facilities use the Minimum Data Set (MDS) for all patients on admission to the facility and at regular intervals throughout their stay. The comprehensive, standardized assessments would promote better quality of care to individual residents by facilitating problem identification and thereby improving care planning. Medicaid nursing facilities in select states have used data from the MDS assessments to specify prospective payment categories for patients. Since July 1998, CMS has required that facilities submit residents' computerized MDS assessments to state and national MDS repositories, thus making these data available for regulatory and comparative purposes.

Presently, nursing facilities conduct quality assurance programs of varying degrees of intensity and sophistication. Some assign personnel to audit charts for information the facilities deem relevant to providing quality of care (e.g., development of pressure ulcers). Other facilities use computerized MDS information to monitor quality. The latter group has the advantage of minimizing additional data collection requirements while examining more complex indicators of quality (e.g., change in status indicators). Because they are uniform and nearly all facilities use them, MDS-based QIs facilitate comparisons across facilities. Certain vendors, called “ORYX” vendors, are Joint Commission on the Accreditation of Healthcare Organizations (JCAHO)-approved; they use computerized MDS information in monitoring the performance of facilities. These vendors also supply JCAHO-accredited facilities with performance indicator information. Contracting with ORYX vendors is a requirement for facility accreditation by JCAHO.

Work in the area of nursing facility quality of care also exists that uses information going beyond the MDS. Dr. Charlene Harrington at the University of California, San Francisco, has examined the relationship between Online Survey Certification and Reporting System (OSCAR) data and staffing standards in nursing facilities. This work (Harrington et al., January 1999 and Harrington et al., April 1999) has led to recommendations for minimum staffing standards. Dr. Andrew M. Kramer and associates, of the Center on Aging at the University of Colorado Health Sciences Center, have developed 80 QIs that use a combination of data from the MDS, nursing facility records, staff interviews, and resident observations and interviews (Kramer, 1999). Under another CMS-sponsored contract, Rosalie Kane and others of the University of Minnesota will develop and test quality of life QIs.

These hybrid QIs combine new data collection with existing MDS data and thus were not compatible with the present task. However, the works of Harrington, Kramer, and Kane will help inform later tasks of the present project.

MDS-based QIs are central to CMS's revised survey process. Facilities that perform well, as indicated through QI rankings, may undergo inspection less frequently than those with greater problems. CMS has instructed surveyors to use 24 QIs developed by CHSRA to guide their on-site surveys and to determine the frequency of surveys. The project team has empirically tested the CHSRA QIs and presents the results in this report.

While there is some documentation of the measurement properties (e.g., risk adjustment) of MDS-based QIs and their utility in accurately identifying patients' problems, researchers have published little work on how well existing QIs perform as measures of facility quality. Required analytical research includes examining the adequacy of the risk adjustment methods used, as well as determining the QIs' performance consistency across states and over time periods. Facilities should be aware of which QIs are reliable and valid when comparing performance across nursing units or over time. Moreover, government regulatory processes and public reporting of facility performance both demand QIs with rigorous measurement properties. CMS and other potential users need to know which QIs are ready for use and the limitations of these QIs.

The present report represents the first stage in developing and validating QIs for post-acute and long-term care (LTC) settings; it outlines the steps taken, presents the results of the empirical analyses, and discusses possible interpretations of the findings. It is beyond the scope of this report to explore the construct of quality of care and identify all relevant dimensions of quality. Future work of this project will further define quality to determine the extent of measurement of all relevant domains and to ensure that domains are neither over- nor under-represented. Evaluation of existing measures, which is the focus of this report, has helped define a framework for evaluation applicable to future validation efforts. The process of review and analysis of existing QIs has also helped guide the project team's procedures for developing new measures.

1.1 Methodological Approach

There were four distinct steps or phases in this evaluation of existing QIs: 1) the search for existing QIs; 2) review; 3) selection of QIs for empirical analysis; and 4) empirical analysis using MDS databases from five states. The project team directing this study consisted of principal investigators from Abt Associates (Terry Moore), Brown University (Vincent Mor and Katherine Berg), the Hebrew Rehabilitation Center for the Aged (John Morris and Kathy Murphy), and CMS (Sue Nonemaker). Four substantive advisory teams, consisting of clinicians and other experts from the field, provided preliminary review of all selected QIs. Programmer-analysts supported the advisory teams and implemented the statistical analysis. The project team was responsible for final review and recommendation of QIs for immediate use. All work presented here was reviewed by a Technical Expert Panel. The four steps are enumerated below.

Search for existing QIs: The project team used a variety of sources to search for QIs. Sources included published literature accessible in *MEDLINE* (medical research), *PsychLit* (psychological research), *Sociological Abstracts* and *Sociofile* (sociological research), and *ERIC* (educational resources), as well as information from the Internet and specific industry and research sources. Industry sources included JCAHO-approved ORYX vendors, nursing facility chains, and nursing organizations.

Review: QIs sent for review comprised all QIs passing a cursory check for relevance and having any possibility of being operationally defined with MDS data. Those reviewing the QIs were advisory team members, including clinicians with expertise in the QI area. Reviewers received all published articles pertinent to a specific area such as pressure ulcers, incontinence, or functional decline. These articles often contained more than one QI, and various sources may have used the same QI. Reviewers made the determination of which QIs represented unique markers. Project staff then entered each different QI as one record, but up to five different sources could reference each QI. Project staff entered ORYX vendor information that arrived after initial allocations had been made, entering a total of 143 QIs into an ORACLE database.

Selection Process: If a QI was operationally definable using MDS data, the project team selected it for empirical analysis. Also requiring analysis were all the QIs CMS had selected a priori for its survey process, even though they were not risk-adjusted.

Empirical analyses: The project team's empirical analyses used longitudinal MDS+ V1.0 data from five states. The analyses examined three aspects of the data: the distribution of facility- and resident-level data within and across all states represented in the database, the consistency of the findings across states and across time periods, the relationship between raw and adjusted QI rates, and the interrelationships among QIs. The team also implemented additional analyses of right censoring (i.e., differential discharge of patients for reasons such as death or transfer) and ascertainment bias (i.e., a class of measurement error reflecting differences in the comprehensiveness and intensity of facility assessment practices that result in the under-identification of resident clinical problems such as pain or depressive symptoms). At the end of this process, project team members selected the QIs for nursing facilities' internal quality improvement programs, and the QIs for government regulatory surveys.

Having decided which of the existing QIs to recommend, the project team refined their risk adjustment method and constructed the refined QIs using MDS V2.0 data from Massachusetts. The results from theses analyses then served as basis for final recommendations by the project's Steering Committee.

1.2 Organization of the Report

Chapter 1 presents the overview of the project and methods, and presents a road map for the organization of the report. Chapter 2 discusses the desired measurement characteristics of QIs and the statistical issues involved in using QIs to evaluate the performance of facilities. Chapter 3 describes the search methods and results. Chapter 4 describes the process of selecting QIs for empirical analysis. Chapter 5 describes the databases used for validation and outlines the analytic plan. Chapter 6 presents findings and a description of all QIs evaluated. Chapter 7 describes a new approach to improve the existing QIs by adjusting for facility-level characteristics and provides final ratings of the selected QIs by the Steering Committee. Finally, Chapter 8 contains a summary of the findings and cautionary notes regarding the use of QIs. Supplemental technical chapters are included in the appendix.

2.0 Conceptual and Technical Issues in Evaluating the Quality of QIs: General Background

2.1 Overview

Health care QIs are a type of performance measure, and as such their lineage dates back to industrial process control systems literature. Industrial applications of these techniques seek to reduce the variance between ideal and actual performance. Data monitoring helps pinpoint the production process area or areas contributing the most to variance (errors) in performance. Little ambiguity exists in industrial applications about the optimization goal, minimizing errors at the highest efficiency.

Over the past decade, the service sector has embraced process control techniques such as continuous quality improvement. However, in the health care field, ambiguity and complexity inherent in the definition of quality complicate the measurement of processes and outcomes. This is particularly true in long-term care, which has goals of care not limited to a specific, definable treatment objective. For the long-term care patient in a nursing facility setting, the aim is not simply to minimize the duration of treatment nor to expedite a return to some prior level of functioning. In fact, curative goals are often secondary to goals related to the prevention of decline. Thus, this arena tends to have many possible goals, not all of which are compatible. For instance, in the long-term management of chronic disease, some have proposed that quality of life and autonomy may conflict with quality of care.

Information about quality of care has several possible constituents and many potential uses. The nursing facility needs aggregated information regarding the quality of care provided and the resident outcomes experienced, both to target efforts to improve the care rendered, and to deliver that care at a reasonable price. The regulator needs this information to target on-site inspections and quality monitoring processes, and to document instances of observed deterioration in care provided. The intermediate purchaser of care such as Medicare, Medicaid, or even a managed care insurer, might use the results of quality monitoring systems to contract with the "best" provider for their beneficiaries and enrollees, or at least to avoid contracting with poor providers. Finally, the consumer and his or her advocates want information on quality both to guide selection of a long-term care provider and to focus political pressure for system-wide improvements in care or reimbursement rates.

Consistent with these different goals and constituents, particular approaches to the measurement of quality can vary in a number of ways (Donabedian, 1966; IOM, 1990; and Blumenthal, 1996). Assessments of quality may focus on individuals using care or on those providing care; their use may be to rate provider performance (e.g., by judging it acceptable or unacceptable, or better or worse than for a comparable organization) or to improve provider performance (e.g., by linking outcomes to processes of care), or both. Internal assessments (i.e., by those providing or directly supervising care) and external assessments (i.e., by regulators, accreditors, or purchasers) both yield important information.

Assessments' rating may be implicit (i.e., by a physician or someone else without current reference to defined written standards) or explicit (i.e., based on written criteria).

In the acute and ambulatory care sectors, performance measures based on processes of care predominate (e.g., immunization rates for children, and mammography and other preventive screening tests among adult women). Increasingly, providers incorporate selected and quite specific patient outcomes, such as glycemic control among diabetics, as measures of quality (Hofer et al., 1999). Some managed care organizations (MCOs) monitor the performance of their physician groups, and the National Committee on Quality Assurance (NCQA) requires MCOs to report aggregated data about these physicians across all participating providers. NCQA then summarizes this information as ratings which it can then distribute to intermediate purchasers of health insurance and to consumers.

While these goals for improving quality are laudable, the need for caution remains. For example, recent revelations concerning the quality of the data underpinning all these MCO performance measures, as well as increased concern about the statistical treatment of the data, have caused some researchers to question the value of provider profiling (Bindman, 1999). To proceed to build a quality monitoring system on questionable inputs, with under-specified control models, is a risk to avoid if at all possible.

2.1.1 The CMS Long-Term Care Quality Indicator Initiative

CMS's new data-based vision for long-term care facility quality monitoring depends on a variety of conceptual and technical assumptions about the validity of QIs and the data defining them. Since no system can ever be perfect, and since technical deficiencies are bound to exist, it is crucial to scrutinize the foundation defining the QI system. In this project, CMS established that charge, and the contractor project team has directly addressed many of these issues. The project team has also established criteria for reviewing the adequacy of existing QIs and for reviewing the QIs' readiness for application to various audiences; these criteria can also help establish the long-term agenda of this research. An appreciation of these complex issues, balanced by an understanding of CMS's required time frame for implementing QIs in the long-term care sector, informed the project team's initial reviews, selections, empirical examinations, and final recommendations.

For a QI system to work in practice, it must be responsive to the following technical issues:

- the availability of a system of data documentation, cleaning, and storage, that will generate the information foundation of the QIs;
- the clinical meaningfulness of the concepts being measured;
- the qualities, including reliability and validity, of the QIs' data elements;
- the specification of meaningful QIs; and
- the specification of adjustment models for the QIs.

In addressing these issues, the project team first summarized the literature on the reliability, validity, and responsiveness of the core of virtually all QIs in current use in the U.S. long-term care industry, the resident assessment data. Here, the nursing facility sector has an enormous advantage over the ambulatory and the acute care sectors, since nursing facilities electronically collect relatively detailed, clinically meaningful resident-level data on an ongoing basis for all residents. However, if aggregates of these resident-level data are to be useful as indicators of the quality performance of facilities, the basic data must be accurate. Various conceptual and technical issues can complicate the construction of aggregated measures of quality, so the project team addressed them, providing several examples based on current experience with QIs in demonstration programs. Finally, the project team summarized the relative importance given these conceptual and technical issues in selecting and evaluating existing QIs.

2.2 Measurement Properties of the Data Used in Creating QIs

Reliability, validity, and responsiveness are three measurement properties essential to evaluating the data used in constructing QIs. In the original development and subsequent retesting of the Resident Assessment Instrument (RAI), the basis for the MDS, researchers undertook a number of studies to establish the soundness of the RAI instrument from a measurement perspective. Subsequent analyses of the resulting data, measurement successes, and to a more limited extent the ongoing problems with the RAI, have all been reported in the literature. Understanding these findings means recognizing that the MDS measurement system addresses a wide spectrum of domains. The MDS does not just look at one, two, or a few problems; it looks at many problems. Its approach is broad, with well-specified items and response alternatives, as well as in many cases, series of items concerning specific aspects of functioning. The MDS approach is to measure functioning, i.e., what the resident does for himself or herself.

2.2.1 Reliability of the MDS Instrument

Reliability refers to the consistency or repeatability of ratings produced by different individuals as well as to the internal consistency of the data elements requisite in measuring a uniform concept. Appendix 1 provides the inter-rater reliability estimates for individual MDS Version 2.0 items that form the basis of many QIs. The average reliability estimates indicate high agreement when 2 clinicians independently assessed the same patient during the same time period. In most cases these inter-rater reliability estimates represent significant improvements over those achieved in Version 1.0 of the MDS and approach or in many instances exceed the reputed reliability of similar tools in research applications (Morris et al., 1997).

Researchers repeatedly tested the inter-rater reliability of trained nurses who used MDS during its initial development (Morris et al., 1990; Hawes et al., 1995; Morris et al., 1997). An inter-rater reliability study of an MDS field test conducted in 1989 and 1990 included nearly 123 residents in 13 facilities and revealed high levels of agreement on the vast majority of MDS items. Inter-rater reliability levels were lower for assessments performed on residents with serious cognitive impairment, indicating the importance of effectively communicating with residents (Phillips et al., 1993), and the simple reality that without the possibility of verbal communication, or sufficiently accurate verbal communication, assessment is dependent on observation. The reliability of some MDS items, such as Activities of Daily Living, or ADLs, will not be affected; in other areas such as pain and mood, where patient self-report is crucial, lower reliabilities are probably unavoidable since the level of clinical discretion and training required to elicit behaviors indicative of the presence of pain or depression is considerable. Morris and colleagues performed extensive testing of MDS Version 2, the revised version of the MDS, with all pre-existing items that had achieved acceptable levels of inter-rater reliability, and found that all new items replacing earlier versions achieved significantly superior levels of reliability (Morris et al., 1997).

Another study compared the accuracy of 23 items of data (using the research nurse as the "gold standard") before and after the introduction of MDS and found the accuracy of patient records increased significantly with the advent of the MDS (Hawes, 1997). More recent efforts to modify and expand the MDS to the post-acute population of patients discharged from hospitals have also included new reliability trials, with similarly positive results in a diverse population of volunteer settings.

In addition to documenting inter-rater reliability, researchers have constructed a number of multi-item summary indices, described below, from items in the RAI. These indices' generally strong inter-item consistency has been sufficiently great for their use as single summary measures characterizing residents' functioning, mood, or psychosocial well-being.

2.2.2 Validity of the MDS Instrument

Validity refers to whether an item or scale actually measures what it purports to measure. As gold standards are rare, validation generally involves accumulation of information from a variety of sources and studies. Several studies support the content, criterion, and construct validity of MDS items and subscales, and despite some variations, the results are very encouraging.

Researchers have compared individual MDS items, as well as summary scales of MDS items, to standardized research instruments used by expert clinicians. The results generally reveal that at facilities with nursing staff trained in the use of the MDS, the MDS items are strongly correlated with research instruments designed to measure similar constructs. For example, Morris and colleagues constructed a cognitive performance scale (CPS) from MDS items and found strong correlations to the Mini-Mental State Exam (Folstein et al., 1975) and to the Test for Severe Impairment (Albert and Cohen, 1992 and Morris et al., 1994). With different samples and somewhat different tests, others have shown similarly high levels of validity for the mental status items in the MDS (Hartemaier et al., 1994). Frederiksen, Tariot, and De Jonghe (1996) found that individual MDS functional status, cognitive impairment, and communication measures have high correlations with comparable research rating scale scores. More recently, Morris and his colleagues have shown that in the area of pain assessment, the MDS Version 2.0 pain frequency and intensity items, when combined in a scale, had high correlations with the traditional research measurement of the pain visual analog scale (Fries et al., in press). On the other hand, the Version 1.0 MDS mood items did not correlate well with research mood scale scores, although crosswalks using the Version 2.0 MDS mood item had very different and positive results with two outside measures (Burrows, 2000).

Several studies examined the relationship between MDS items on continence and the volume of urine measured among intermittently incontinent residents in nine nursing facilities (Crooks, et al., 1995). In one study, the relationship between the MDS items and measured volume was significant but weak, largely due to the fact that in several facilities there was virtually no relationship between the two measures, whereas other facilities manifested a highly positive relationship between MDS-rated incontinence and urinary volume. In another MDS-based incontinence study, the results were much more positive (Brandeis et al., 1998).

In another area, a comparison of the MDS item summarizing vision and formalized clinical tests of visual acuity found agreement in only 34 percent of residents, and found that the MDS missed visual defects that were clinically observed to be present (Swanson and Glick, 1995).

A recent series of studies using MDS data from five states participating in CMS's Nursing Home Casemix and Quality demonstration confirms the validity of the diagnostic and functional outcome data in the MDS (Gambassi et al., 1998; Bernabei et al., 1998; and Landi et al., 1998). By comparing the diagnoses in the MDS with diagnoses on Medicare hospital claims and the types of medications taken, the authors found high levels of internal consistency and congruence between these data sources. Additionally, a comparison of mortality, hospitalization, and functional decline rates of congestive heart failure patients taking ACE inhibitors or digoxin, based on MDS data, replicated the results of numerous randomized studies, and demonstrated the validity of the data (Gambassi et al., 1998; Bernabei and Gambassi, 1998).

2.2.3 Responsiveness of the MDS Instrument

An additional measurement property of the MDS is responsiveness to change. For clinical or research measures, responsiveness refers to the ability to detect clinically relevant change, even change of small magnitude. This measurement property is essential in monitoring patients' status over time and in evaluating treatment effectiveness. It is also important that QIs display real rate differences in response to quality improvement interventions by facilities. No research had assessed any of the existing long-term care QIs for responsiveness to change prior to this project team’s analysis.

Several studies have examined the ability of the MDS items to track resident functional status, and have constructed summary scales to measure changes in resident functioning over time. As part of the MDS evaluation project, Phillips and his colleagues found that measures of physical, cognitive, emotional, and social functioning, as well as individual data items, all revealed substantial sensitivity to change over six months (Phillips et al., 1997a). MDS measurements such as functional status, cognitive status, and clinical severity have also shown predictive ability for future hospitalizations and mortality (Mor et al., 1997). Gambassi and his colleagues (1998) found that changes in functional status were sensitive to exposure to selected classes of cardiovascular disease drugs. Morris and his colleagues (Morris, Fries, and Morris, 1999) have demonstrated the utility of the MDS functional measures in detecting ADL changes for patients receiving exercise therapy or nursing-based rehabilitation. Landi and his colleagues (1999) found that changes in the Cognitive Performance Scale had independent relationships with both mortality and hospitalization.

Recent analyses of decedents also revealed that the closer the last MDS assessment was to the date of a patient's death, the more dramatic was the increase in prevalence of selected symptoms associated with "terminal decline" (Miller and Mor, in press).

This review suggests that the functional and cognitive status constructs included in the MDS are related strongly to research-quality instruments, when residents' records are the sources of the MDS data and research data gathering occurs under controlled conditions. However, not all the items correlate as well as these with established clinical measures.

The accumulated evidence on the reliability and validity of the MDS items and subscales bodes well for the measurement properties of QIs developed from the MDS. However, when these items form the basis of potential QIs, it is important to additionally assess their performance for that purpose. The following section examines a variety of interrelated conceptual and technical issues that can greatly influence the validity and applicability of QIs regardless of how reliable, valid, and responsive the underlying resident-level data are.

2.3 Conceptual and Technical Issues in Aggregating Data into QIs

In the past, researchers have frequently proposed and used measures of nursing facility quality, but generally only for a small number of facilities or in select groups of facilities. Until recently, most such measures have used aggregate data about each facility as their basis, and have compared the rate of "events" between facilities with various characteristics. For example, Zinn and her colleagues used facility-level survey data to test the effect of facility staffing and market factors on indicators of quality of care (Zinn, Aaronson, and Rosko, 1993; Zinn, 1994). Nyman (1988) also relied on aggregate data to examine the effect of different types of facility characteristics on selected QIs. The use of these data limits these studies, in that risk adjustment is quite minimal due to the ecological fallacy, i.e., in an aggregated database no risk adjustment is possible on the patient level. This is one of the reasons that so many of the early studies of the determinants of quality of care in nursing facilities led to contradictory findings (Davis, 1991).

Use of individual-level data to conduct studies of nursing facility quality requires large numbers of nursing facilities and data on large proportions of residents. These data almost always come from administrative sources. One early data source of this type, used in a number of quality-related studies, is the National HealthCorp (NHC). Located primarily in South Central U.S., this is a chain of approximately 100 facilities that has maintained longitudinal, resident-level assessment data on all patients since the mid-1980s. Many researchers have conducted studies using these data, including studies investigating organizational determinants of quality of care; the impact of Medicare policies; and the clinical determinants of falls, pressure ulcers, and hospitalization (Kiel et al., 1991; Mor et al., 1993; Morris et al., 1994; Brandeis et al., 1995; Ooi et al., 1999). Medicaid data, particularly available hospitalization and medication use information, have also been a good source of patient-specific information for creating indicators of the quality of care (Lipowski, 1996). Outside the United States, Shapiro and Tate have linked Manitoba, Canada, nursing home patient data to administrative data, examining the effect of organizational factors on the length of time between a resident's admission and the resident's experience of a negative outcome (Shapiro and Tate, 1995).

The introduction of the RAI into U.S. nursing facilities significantly altered the care quality arena. Four key events in the history of this system are relevant: the national implementation of MDS Version 1.0 in 1991; the introduction of MDS Version 2.0 in 1996; the state-mandated computerization of MDS Version 1.0 in over 10 states in the early 1990s (and thus the availability of an MDS database with which to create the QIs this report analyzes); and the national mandate in June 1998 for full MDS computerization in all nursing facilities, with the associated introduction of state and national repositories for these MDS computerized data. With these resources, planners began to describe a system for designing, testing, and widely disseminating QI systems for nursing facility-based care.

The MDS measures the functioning of an individual patient; QIs can be measures which characterize the performance of the nursing facility, when aggregated from the resident level to the facility level. At the resident level, researchers must identify meaningful problems that will be the subjects of the quality measures. The resulting system must also define the resident-level measurement and covariate models to explain those phenomena. Using data gathered from facilities and in states, researchers next have to establish how the models apply across these various settings and environments. All of these measurement models must be aggregated to the facility level, and inter-facility QI rates compared. Considerations of the reliability, validity, and responsiveness of these facility measures represent special cases of general measurement model issues, and are complicated by the fact that researchers must continually focus both on the resident as the unit of measurement and on the facility as the unit of analysis.

2.3.1 Content Validity and Multi-Dimensionality of QIs

Since this report focuses on examination of existing QIs, it was not within the purview of the project team to address the complex issue of defining quality. Rather, the team's starting proposition was that the basic concept measured by an existing QI does relate to quality. Review of existing QIs revealed that they covered numerous dimensions, ranging from specific clinical phenomenon such as dehydration, to generalized notions such as psychosocial well-being. To span all these categories, the classification schema placed QIs into the following domains: functional; clinical; psychosocial; and pharmacotherapeutic. Theoretically, the project team anticipated that the QIs within each domain would be correlated, particularly those QIs using the same measurement construct.

Classification of QIs into a single measurement domain was necessarily arbitrary. For example, the project team could readily have classified depression without drug treatment as pertaining to pharmacotherapy rather than to the psychosocial domain. With that in mind, the project team reasonably based their decisions on very specific characteristics of each QI relating to the QI's basic content validity. For example, a QI predicated on the use of a functional measure of mobility might not really measure functioning, but rather facility rules or practices related to use of mobility aides (e.g., wheelchairs). If the underlying concept of that quality outcome were purportedly to measure functional independence but the data actually combined wheeling and walking, the QI might measure another concept altogether. Thus, crosswalking between the measurement and meaning of the individual MDS data element and its interpretation (or misinterpretation) at the aggregated level was a crucial exercise for the teams reviewing each existing QI.

Quality of care is necessarily multidimensional, meaning that no single QI is likely to capture overall facility quality. Facilities may perform extremely well on one type of QI, but may not perform nearly as well on another. Indeed, two papers recently confirmed this hypothesis, one using New York state data and the other data from Massachusetts (Mukamel and Brower, 1998; Porell and Caro, 1998). Thus, the creation of a single aggregated measure or "picture" of facility performance is not a simple undertaking. Rather, creating a single quality metric for rating nursing facilities is unlikely.

2.3.2 Operational Definitions

The most basic criterion for evaluating existing QIs was the clarity of the operational definition. The components of a QI are the numerator, the denominator, and, if present, factors used to adjust for, or stratify, the casemix within facilities. The numerator refers to the upper portion of a fraction used to calculate a rate, proportion, or ratio, e.g., patients who have the characteristic or outcome of interest. The denominator is the lower part of the fraction, and may refer to those at risk of developing the outcome or characteristic of interest, or may refer to all persons in the facility.

An exact definition of a QI requires: precise instructions regarding which data elements constitute the measure, the methodology for aggregating the data, the data element combination of each composite measure, and any further instructions regarding stratification or risk adjustment. Without clear instructions and a rationale for these choices, it becomes difficult to know whether the QI proposed by the developers is being correctly replicated.

Translating from one version of the MDS to another for testing or for interpretation added an additional complication to the task of reviewing and applying existing QIs. Furthermore, some QIs had actually been developed using data that predated the RAI. Testing QI measures from New York's PRI tool, for instance, required that the project team translate that QI's earlier instructions into an MDS version. Ultimately, lack of additional clarifying statements from the originators of the QIs reviewed led to the inability to empirically evaluate some selected QIs.

2.3.3 Documenting Rare or Sentinel Events

Traditionally, measuring quality means accumulating information about the rate at which clinically undesirable and avoidable events occur. All hospitals and most nursing facilities have reporting systems in place for selected classes of "events"; these may range from falls, to adverse drug reactions, to infections. These events may have drastic consequences for the particular resident (e.g., falls), or for other residents (e.g., infections). Facilities monitor them precisely because of these potential consequences, and because many people believe that these rare events can be minimized or prevented altogether.

Without specialized reporting systems, however, facilities cannot monitor transitory and rare events on an ongoing basis, and so cannot use such events in measures of quality. A quarterly measurement system such as the RAI cannot precisely measure episodically occurring events, nor detect rapidly resolving clinical conditions such as fever. A high temperature can have a rapid onset, be subject to aggressive pharmacological treatment, and resolve in a matter of days or weeks, not months. Therefore, use of the MDS to monitor a patient for fever every 90 days, misses a significant number of cases in which patients experience fever in the period between assessments. A sentinel indicator of quality could be based upon a rare event if it had a longer time frame than the transitory events; it would be more amenable for use as an indicator of quality deterioration.

The concept of a sentinel event QI is one that suggests that the presence of a "signal" episode might signify the presence of a quality problem. New falls, new pressure ulcers, or new restraints on residents who did not have them at the prior MDS assessment are candidates for this type of QI. An observation that one out of 96 residents in the facility without a stage 3 or 4 pressure ulcer at the prior assessment had an ulcer at the most recent assessment could be such a sentinel QI, once again recognizing the possibility of missing some patients who might have acquired and resolved a pressure ulcer during the interval between assessments.

In facilities serving a similar casemix of residents, those facilities in which certain rare events are more prevalent might have more of a care problem than other facilities in which such rare events are virtually non-existent. Obviously, where the QI is less rare in the population and the true denominator is sufficiently large, and the QI is risk-adjusted, more confidence is possible that a facility with a higher than expected rate of the problem actually has a quality problem. However, even in these cases, the QI is still only that, an indicator, a probabilistic estimate of the likelihood that the facility has a real quality problem in the given domain of care.

These measurement issues are one reason these empirical analyses of existing QIs followed the rule that a facility must have at least 20 residents in the denominator in order to generate a usable score on a QI. Based on simple plots of the relationship between the reliability of an estimate and the number of residents in the facility, it was clear that unless the facility effect was extremely large, facility-level QIs calculated on fewer than 20 residents would not be reliable.

2.3.4 Ascertainment Bias

Ascertainment bias is a special class of systematic measurement “error”. Ascertainment bias occurs because of variation in the ability and attentiveness of nursing facility staff in conducting an assessment of resident status, particularly in domains that are more difficult to delineate, such as delirium, mood distress, pain, or dehydration. This means that staff in some facilities have the necessary skills in, and are more attentive in, monitoring and noting the presence of clinical problems, for which intervention strategies can then be implemented. On the face of it, a simple comparison of facilities would reveal that such facilities appear to have higher than normal proportions of residents with these conditions. The key phrase is "appear to have ", for the differentiating factor is not one of problematic care but rather of assessment acumen. Facilities with more diligent clinical assessors might therefore receive a rank indicating worse care than other facilities, though from an objective viewpoint the problem might be similarly prevalent in the two homes, with the staff in one of the facilities being less likely to identify or record the problem. Under-identification of clinical problems in such facilities may be due to purposeful coding policies, but it is more likely either that

staff members have "normalized" these types of clinical problems and do not record them as remarkable, or that they lack the skills to identify affected patients. It is not unlikely that facilities with better-trained staff and superior practice patterns are more diligent in their recording of resident symptoms. Thus, differences in assessment can lead to systematic distortions and bias QIs against facilities with good quality of care.
Signs of the presence of this type of misclassification of the basic RAI measurement are that some facilities will have much higher, or much lower, rates of a particular clinical syndrome than most other facilities, even after taking into consideration the different types of patients in the facilities. Better understanding of the scope of this issue required an extensive series of empirical analyses to examine how much this form of resident-level measurement bias might, in turn, bias the facility measures of performance (see Technical Appendix 1).

2.3.5 Heterogeneity of the Population and Risk Adjustment

Recent increases in the types of services offered by nursing facilities have led to greater differentiation between facilities and have placed the industry as a whole in a much more competitive position relative to an array of other health care providers. For example, over the past seven years the numbers of facilities with Alzheimer's disease special care units has more than doubled. In less than 10 years there has been a nearly 10-fold increase in the proportion of total nursing facility beds housing Medicare-reimbursed patients. The categorization of these changes can either be as service diversification in response to demand from specific subpopulations, or as specialization to capture a larger share of an already existing sub-population. Mor, Banaszak-Holl, and Zinn (1996) reported that the trend toward diversification increased over time across a number of different special care services. Castle, Mor and Banaszak-Holl (1997) found that facilities that already had a special care unit were more likely to establish a hospice special care unit. All of these signs point to increasing heterogeneity in the nursing facility industry and increased specialization across facilities. This means that increasingly, the types of residents served in one facility are not necessarily comparable to those served in another.

Casemix differences in the types of residents living in facilities make direct comparisons of the rate of various clinical problems in those facilities difficult. Research by Zimmerman and his colleagues (Arling et al., 1997) as well as Morris and his colleagues (Ooi et al., 1999; Mor et al., 1997; Brandeis, et al, 1995) strongly point to the need to risk-adjust most outcomes in order to make adequate comparisons between facilities. In light of the fact that different facilities may attract different types of residents and that the risk of having a negative care event may vary substantially as a function of resident characteristics, the need for some forms of risk adjustment becomes very important. One type of risk adjustment, traditionally used implicitly, is the differentiation between long- and short-stay residents. Facilities with many new admissions per quarter will have patient populations whose clinical problems may reflect conditions prevalent in acute hospitals. Maxwell and her colleagues (1998) applied the QIs developed by Zimmerman and his colleagues under CMS's Nursing Home Casemix and Quality demonstration to new admissions and long-stay residents and found very different rates across many measures.

Most of the QIs identified and evaluated in this report implicitly or explicitly addressed the long-stay resident either by excluding consideration of any data from new admission assessments or by requiring that the resident have two successive assessments. This focus on longer-stay residents contributes to reducing some of the most complicated heterogeneity in the nursing facility population and, as such, represents an important first step toward equating the populations. Nonetheless, even among long-stay residents, it is clear that facility casemix matters in facility comparisons; e.g., facilities specializing in Alzheimer's disease cases will differ in many measurable and unmeasurable ways from those serving specialized, medically complex cases. While adjustment for such differences in resident populations using statistical or stratification models is important, the impact of real differences in casemix on observed outcomes and the associated measured QIs can never be assured.

2.3.6 Facility-level Effects in Resident-level Risk Adjustment Models

The issue of risk-adjustment is problematic in the long-term care industry. Many nursing facilities care for a wide range of acuity levels as they attempt to balance the provision of post-acute care with traditional long-term, chronic care. Because of the unique position of nursing facilities along the continuum of care, risk-adjustment must go beyond the resident-level to include some measure of the profile of residents admitted. For example, in the case of pressure ulcers, facilities care for residents with varying levels of intrinsic risk (e.g., functional problems, diabetes, incontinence). Yet nursing facilities also inherit the results of the good and poor care practices of hospitals. The poor practices of hospitals are readily apparent in the profiles of nursing facility residents at admission. Based upon all MDS admission assessments in New York in 1999, the average rate of pressure ulcers recorded upon admission to the nursing facility was 18 percent.

Clearly this rate was not the same across all facilities in the state. Some nursing facilities specialize in admitting these more clinically complex discharges from hospitals, while others operate to discourage the admission of such patients. Either way, facilities tend to select patients, or to have them referred, from hospitals that closely match their resources, skills and mission. Since residents with a history of a pressure ulcer are significantly more likely to acquire one in the future (for physiological and even measurement reasons due to difficulties in reverse coding), facilities admitting patients with pre-existing pressure ulcers run the risk of looking worse on a pressure ulcer QI simply because they admit a higher acuity population at admission. This selection phenomenon can undermine the actual and perceived fairness of the QI comparisons. Facilities will have a disincentive to provide care for the most vulnerable if the QIs adopted by the government or accreditation agencies fail to properly adjust for this selection phenomenon. The same principle also operates in other clinical areas.

Selection is closely related to ascertainment bias—precisely because facilities specializing in treating patients with selected problems are likely to do a better job of identifying and measuring pertinent clinical characteristics of those types of residents. For example, nursing facilities admitting a high percentage of patients with pressure ulcers are more likely to identify pressure ulcers at admission and at subsequent quarterly assessments and hence are likely to appear worse simply because they record more problems. Failure to account for facility variation in admission profile may lead to QI flags that are not driven by a problem of quality of care, but due to facility specific admission practices and associated clinical care documentation. In view of the impact these facility-level effects may have on relative quality ranking of facilities, the project team completed a set of analyses that included adjustments for facility-level effects in estimating the expected QI prevalence for a given facility. Chapter 7 summarizes these analyses.

2.3.7 Censoring via Transfer, Hospitalization, and Mortality

Nursing facilities differ with respect to the rate at which they discharge residents, just as they differ in the types of residents they admit. Discharges of long-stay residents can be due to hospitalization, death or even transfer to another nursing facility. Recently published data reveal that facilities vary substantially in terms of the rate of hospitalization of their residents. Facilities that have more skilled nursing and medical staff have lower hospitalization rates (Intrator, Castle, and Mor, 1999). Re-analysis of the MDS evaluation data set collected in 10 states revealed large inter-state variation in hospitalization rates among nursing facility residents in the last six months of their lives (Mor, 1999). Mor and his colleagues (1997) found that inter-facility transfers, even via an intervening hospitalization, are relatively rare (less than 2 percent of all transfers from the nursing facility). A recent study replicated this finding by examining all transfers in New York and Maine (Hirth, Banaszak-Holl, and McCarthy, 1999). Nonetheless, the observed differences in hospitalization rates, and even mortality rates, clearly could bias the measurement of any QI predicated on only the resident population measured using the MDS (Mor et al., 1997).

If some facilities discharge residents who begin to manifest signs of negative outcomes, the true extent of quality problems present and documented in those facilities will be underestimated. Even risk adjustment would be unable to control for this phenomenon since discharged, or "censored", residents will not have a measurement. This type of confounding is very likely if the risk factors are diseases or conditions that predispose residents to manifest a facility quality problem as well as to increase their risk of death or hospitalization, and thus of censorship.

In light of the potential impact this phenomenon may have on the relative quality ranking of one facility versus another, the project team, using a combination of longitudinal data sets, undertook a series of analyses to determine the scope and potential biases associated with censoring. Technical Appendix 1 summarizes these analyses.

2.3.8 Stability of QI Measures

Nursing facility administrators and long-term care surveyors need to know whether a measured change in the rate of the QI, or in the relative ranking of the facility compared to others, reflects a true change in quality or just random fluctuation of QI rates. To be useful as a marker of quality, the QI rates should remain stable across time periods if there has been no actual change in the quality of care provided.

Using data from the quarterly MDS assessments, multiple performance measures can be constructed to characterize the quality of a nursing facility on a quarterly basis. By definition, measures of the prevalence of a specific clinical condition, if calculated only on the long-stay population, are likely to be very stable from quarter to quarter. Measures of incidence, as well as rare sentinel events, are more likely to be unstable from quarter to quarter. From the perspective of the audience using a QI, whether a measure should be more stable or more volatile depends on how it is used. Facilities using QIs in guiding their continuous quality improvement program might want to track the more volatile quarterly rate to determine whether any of the processes of care might be able to reduce the volatility. In contrast, QIs to be used by consumers and purchasers would preferably reflect a more constant state of affairs, since the users would not want to change facilities whenever a QI drops below some level.

In light of the importance of having a stable measure, one that reflects an underlying concept of an aspect of quality of care, rather than a more transitory concept, the project team explicitly used stability as a criterion for comparison in empirical analyses of the selected QIs. Some QI originators actually construct QIs based on an average of multiple quarters of data precisely to achieve a more stable estimate of the quality of a nursing facility relative to other facilities.

2.3.9 Attributing Variation in Quality Outcomes to Provider Performance

Even under the most optimal circumstances (i.e., high prevalence of the outcome of interest, large sample size, and stability from quarter to quarter), certain statistical issues must be considered in interpreting and applying the results of QIs in ranking nursing facilities. Only the variation in any given patient outcome variable, ranging from the incidence of pressure ulcers to the rate of decline in mobility, that is attributable to the facility in which a patient resides, is theoretically under the control of the facility. The "attributable" effect of practice variation is a theoretical construct, since measuring it depends on having "controlled" for all relevant clinical and patient factors, which are never truly known. Nonetheless, it is well established that the more the variation in outcome "attributable" to residence in a facility, the smaller the number of patient observations needed to calculate the rate of the QI, given the same level of reliability of the estimate (Hofer et al., 1999; Bravo and Potvin, 1991). Thus, for example, if the degree of variation attributable to a facility is four percent, some 100 residents must be in the denominator of a QI to reach a facility-specific prevalence estimate that has a reliability of .80. Depending on the estimated size of the facility effect on an outcome of interest, Hofer and his colleagues recently demonstrated that it is possible for the facility effect to be "statistically significant" but not be able to differentiate the QI estimates between two facilities which are virtually at the extremes of the quality rankings (1999). This means that, although there is a real effect, facility sample size and the size of the attributable facility effect are such that it may not be possible to reliably differentiate facilities that are top performers from those at the bottom of a quality distribution.

2.3.10 Summary

Perhaps the most important aspect of the QI concept is precisely that the measure is a signal of the possibility that a problem may exist, rather than being an assurance that the problem is present. All the issues enumerated above represent reasons not to expect that the observed, or even risk-adjusted, predicted rate of a QI will fully reflect the care provided by a long-term care facility.

In the acute, ambulatory, and long-term care arenas to date, to the project team's knowledge no one has fully tested a set of QIs that address all the conceptual and statistical issues raised above. Furthermore, it is unlikely that any existing QIs, including those used by the National Committee on Quality Assurance to rank health plans and hospitals, would be fully acceptable, based on these criteria. Proponents of provider profiling and "report cards" of performance acknowledge the limitations of these techniques but maintain that having some data is superior to having none at all (Epstein, 1995). The project team agrees. While caution is warranted in the implementation of QIs, particularly for those audiences that will be making final purchasing choices based on the data, the data and thus the ability to develop strategies to overcome the technical problems are only likely to improve on application of these measures in the real world.

2.4 Current Experiences in Applying QIs

Over the past half-decade several sources of information have become available regarding the application and uses of QIs in the nursing facility arena. As might be imagined, there have been no carefully conducted impact evaluations. Most of the information derives from case studies as well as from program descriptions. The project team anticipates that over the next several years more and more information about the adaptation and use of various long-term care QIs will be forthcoming from the JCAHO ORYX initiative as well as from other studies undertaken by the nursing facility trade associations. This section comprises current summaries of selected studies and reports produced about the CHSRA as well as the LTCQ QI projects.

2.4.1 Experience with the CHSRA QIs

In the literature review, few developers provided detailed information on the results of or the methods of assessing the performance of their QIs, i.e., the “validation” of their QIs. The majority of ORYX vendors provided only the definition of their QI or QIs, often without detailed instructions. A notable exception is the work of Zimmerman and associates at CHSRA.

The CHSRA validation process included face validity, content validity, criterion validity (in this case, how well the QI compares to state surveyors' findings about the same facility), and predictive validity (whether the presence of a QI predicts a problem with the quality of care). To validate their 31 QIs, CHSRA researchers implemented pilot testing in 1993-94 and primary validation testing in 1994-97 (see Zimmerman et al., 1999; Zimmerman and Karon, 1997; and Zimmerman et al., 1995). They investigated whether there were problems with the accuracy of each QI, evaluated whether each QI would produce false positives or negatives, and determined whether the QIs were valid indicators of the quality of care at the resident and facility levels.

As of September 1997, CHSRA completed validation studies in nine facilities in three states, with a total of 378 resident-level QIs. Facilities were selected based on the prevalence rates of particular QIs and the need to coordinate the validation teams' work with the state surveyors' annual certification visits. At each facility, CHSRA researchers chose four to five QIs with high rates of occurrence and one QI with a low rate of occurrence to determine the generation of false positives or negatives. The validation team preselected twenty residents at each facility for in-depth review, hoping to observe at least five residents with each of the QIs under consideration in that facility. The team also selected an additional five residents on site. Research teams worked in pairs that included at least one registered nurse.

- First, teams verified the accuracy of the data underlying the QI. They evaluated whether the information on each QI report was consistent with the information on the original MDS+ report form, as well as whether evidence on the resident's medical record supported the information on the MDS+ form, or at least did not contradict it. This analysis indicated any systematic programming errors in generating the QI reports, and whether facility staff had made errors in completing the MDS+ forms.

- Next, the teams determined whether the 'triggered' QIs had identified actual quality of care problems at the resident or facility level, either as isolated examples of poor care, or as patterns of poor care across a facility. They assessed the severity or scope of any problems found, as well as the nature or seriousness of those problems, and also looked at the distribution of quality problems across facilities.

- Third, researchers compared each QI's ability to link quality problems in facilities to "performance thresholds"; that is, the facility's relative ranking on that QI compared to other facilities in the state. The researchers included some facilities ranked between the 75th and 89th percentiles for the state, but whenever possible, they included facilities ranked at or above the 90th percentile.

- Finally, the validation teams assigned F-tags (or deficiency citations) for each problem identified by the QI, so that the QIs could be compared with state surveys conducted at the same time.

Overall, CHSRA's research indicated that QIs with high rates of occurrence, selected at high threshold levels, are useful tools for identifying quality of care problems at both the facility and resident levels. The QIs generally had high accuracy ratings, and most identified severe problems for all or some residents. At the facility level, most of the observed quality problems were severe enough to warrant a citation from the survey team, often based both on the scope and seriousness of the problem. The cases in which the QIs led to the identification of facility-level problems were not evenly distributed across facilities, indicating the need for further research into the process of targeting facilities for review. Researchers can draw few conclusions about the QIs at low rates of occurrence, since they only collected a limited amount of information about them.

The CHSRA team identified the need for further research in identifying possible QI-specific thresholds. In addition, the researchers noted that minimal overlap occurred between the specific F-tags assigned by the validation teams and those assigned by state surveyors on site at the same time; however, on looking at the overlap of broader "issues of concern" between the surveyors' and the validation team's findings, they found greater congruence. It was still worrisome that the state surveyors did not identify many of the quality problems found through the QI validation studies. The CHSRA researchers noted a need for further studies to determine the source of difference between the findings.

2.4.2 Experience with the LTCQ© Q-Metrics Rankings

The LTCQ Q-Metrics© rankings are based on aggregated MDS data in which each individual resident's deterioration on a particular combination of MDS items includes adjustment for residents' risk of decline and then aggregation at the level of the facility. Like the CHSRA QIs, the Q-Metrics© system has multiple outcome domains with which it characterizes the performance of facilities. In most instances outcome measurement uses a variety of different summative scales that are composites of several data elements that assess related aspects of functioning. Testing the reliability of such measures requires examining the inter-rater reliability (that is, combining two independent raters' assessments to create a single measure) and determining the degree of internal consistency of the component items. In general, the more internally consistent a measure is, the more stable and reliable it is in statistical analyses. Using research-based MDS data as well as statewide databases for selected LTCQ outcome measures, the Kappa as well as the alpha reliabilities (degrees of internal consistency) reveal a high degree of reliability. Published research of analyses of longitudinal MDS data show that most of these measures are sensitive to change in resident status, with the measures indicating both deterioration as well as improvement, depending on the status change (Phillips et al., 1997b; Gambassi et al., 1998).

Establishing the validity of any measure of quality is necessarily complex since there are no absolute standards against which to compare any measure of interest. Validity is then the process of iteratively improving the measure and determining its reasonableness based on whether it appears to measure the intended aspects of quality. LTCQ reports the following three broad types of evidence of the validity of this system’s measures.

- First, comparison of the casemix-adjusted average rates of decline for several outcomes were compared to the distributions of survey deficiency rankings by state. This comparison shows a strong correspondence between the distribution of ranks from state to state and the average percent of residents whose outcome declined, with New York having the lowest rate of decline (19.2 percent) and Nebraska having the highest (29.7 percent).

- Second, a relationship has been observed between selected facility characteristics (e.g., structural features such as levels of staffing), and other indicators of quality obtained from the annual state surveyors' inspections. LTCQ found reasonable relationships between organizational and staffing input factors and facility rankings on selected outcomes. For example, a high quality ranking on trunk restraints correlated positively with total staff ratio (r =.13), with the ratio of total licensed staff to residents (r =.16), with total beds (r =.18), and with a variety of different measures of having the capacity to provide highly technical care (e.g., tracheostomy or injections). Of considerable interest is the fact that the number of health-related deficiencies observed at the inspection concurrent with the MDS data was correlated -.29 with the quality ranking on restraints. Since the higher the quality ranking, the better the facility, it makes sense that nursing facilities with more deficiencies have lower facility ranks.

- Finally, anecdotal evidence for the validity of the Q-Metrics© rankings is reported based on experience with several quality care consortia in Massachusetts and Rhode

Island. Most of the facilities choosing to participate in the consortia were high quality providers, based on their reputations among their peers. Furthermore, they tended to have the resources at their disposal to invest in special programming and had the administrative sophistication to perceive the potential value of using the MDS data for more than care planning. During consortium meetings, facilities with high rankings on a given domain of quality share their "best practices" with the others. These high-ranking homes made the MDS program a priority, linking it explicitly to other aspects of the homes' clinical program. Participants in the consortia believed that facilities identified as manifesting the best outcomes on the Q-Metrics© measures had innovative and high quality “best practices”. This phenomenon was evident in diverse areas, from activity programming to pressure ulcer prevention and care.

2.5 Summary of Factors Contributing to Assessment of the Quality of QIs

As noted, the project team devoted considerable discussion to the issues addressed in this section of the report. These issues were manifest in the specification of the descriptive factors staff abstracted from the literature, the web search, the review of the material received from ORYX vendors, and the material supplied by nursing facility chains. In selecting QIs for detailed empirical analysis, the project team relied heavily on the extent to which the various QIs addressed these conceptual and technical issues. Finally, the project team constructed and utilized a guide for selecting which QIs to recommend for further implementation by CMS, based on the perceived acceptability of a QI for a particular audience. Table 2.5 below summarizes this guide. The recommendations made by the project team regarding the use of 22 QIs for different audiences are summarized in Chapter 7.

Table 2.5

Guide for Determining Acceptability of a QI for a Particular Audience

Focus of Measurement	**NURSING FACILITY: Internal QI monitoring, benchmarking**	**SURVEYOR: External QI – drives survey process, accountability, benchmarking**	**PUBLIC REPORTING: Communicating QIs to consumers and purchasers**
Consistency of QI over time periods (quarters)	0	+	+
Potential for censoring bias	0	+ +	+ +
Potential for selection bias	0	+	+
Risk adjustment	+	+	++
Face/construct validity of the QI components	+	+	+
Reliability of variables scales used in the QI	+	+ +	+ +
Degree of potential control by facility over the outcome	+	+	0
Consistency of QI over multiple states	0	+	++
Importance and relevance of the QI (i.e., the “So What” Test)	+	+	++

Note: 0, + and ++ indicate the level of importance of each particular item.

3.0 Description of Search for QIs

This project cast a broad net to compile a database of QIs or outcome measurement systems that could be applied to the assessment of quality of care in nursing facilities. The project team searched a variety of electronic databases, sent letters to all JCAHO-approved ORYX vendors and to researchers and organizations actively working in the area, and searched the Internet for relevant material. The following sections describe the methods and results of the search.

3.1 Published Literature

The project team searched literature published between 1966 and 1999 using the following electronic reference databases: MEDLINE (medical research); PsychLit (psychological research); Sociological Abstracts, or Sociofile (sociological research); and ERIC (Educational Resources Information Center). Using the following combinations of search terms: "QIs", "quality measures", "outcome QIs", and "outcome measures", each paired with "nursing facilities", "post-acute", "sub-acute", "long-term care", and "rehabilitation", the project team tailored the search to include primarily those articles pertaining to long-term care, post-acute care, or special populations. This yielded 37 possible articles.

From the list of 37 articles, the project team forwarded 16 for review, spanning the years 1993-1998. The 16 articles described nursing facility QIs that could be reviewed and evaluated for content related to the scope of the investigation. Specifically, the project team selected QIs that could be evaluated based on the type (domain and representative population), data source (e.g., MDS, CMS claims, internal records), psychometric properties, number and size of nursing facilities using the QIs, and their application and uses. The project team excluded any articles that did not contain specific QIs, as well as articles containing QIs that were irrelevant to this study, such as those for acute care and pediatrics. From the 16 articles, the project team identified and reviewed 57 nursing facility QIs. (See Table 3.1 at the end of this chapter for a summary of QIs obtained from each type of source.)

3.2 Internet Search

The project team also implemented an Internet search to identify organizations with QI systems. From mid-December 1998 to mid-January 1999 the project team used a multi-search engine from Agents Technologies Corp., *Copernic 98*, to search 10 engines simultaneously, documenting records of hits within each search engine. The available search engines within the *Copernic* system were: *Yahoo!, Webcrawler, Magellan, Lycos, Looksmart, Hot Bot, Excite, AOL, and Alta Vista.* Each engine had a 300-match limit and combined search engines had a 1000-match limit using its customized search option, the least limiting of search types.

Using multiple search engines on the Internet, the number of hits does not have the same relevance as for literature searches, because of duplication. To capture selected key theme words, the project team used twelve search-term combinations, including: quality measures and nursing facility; quality measures and rehabilitation; quality measures and sub-acute care; QIs and nursing facility; QIs and rehabilitation; QIs and sub-acute care; quality outcomes and nursing facility; quality outcomes and rehabilitation; and quality outcomes and sub-acute care. The project team also tested the various search capacities and delimiting functions in *Copernic 98* with other combinations, and found that the above key words produced the most exhaustive matches.

Another target for review were the web sites of agencies, companies, clinical institutions, or organizations that had developed, or were presently developing, QIs for long-term care, rehabilitation, or sub-acute populations. Among the targeted web sites were: sites for pure research articles; sites related to Zimmerman QIs; the University of Wisconsin-Madison web site; the Nursing Home Casemix and Quality demonstration web site; the JCAHO web site; sites for ORYX providers for long-term care; educational program web pages; provider home pages; newsletter web sites; newspaper web sites; hospital web sites; sites for agencies on aging; and sites for quality programs (i.e., provider articles). The project team forwarded for further action only web sites that contained specific references or descriptions of QIs, and that were known not to be duplicative to the other search.

Through this effort, the project team identified nine organizations; however, five had already been contacted through the ORYX or nurse executives mailings (see Section 3.3). The remaining four organizations received letters requesting QI information. One organization, the Arizona Healthcare Cost Containment System, responded and provided information on its quality monitoring system; however, because of it lacked specific QIs, the project team did not forward this system for review.

3.3 Correspondence with ORYX Vendors, National Associations, Other Industry Sources

3.3.1 Selection of ORYX Vendors to Contact

A listing of JCAHO-approved ORYX software vendors was available on the World Wide Web at *http://www.jcaho.org/perfmeas/oryx/mtr_frm.htm*. Each vendor listed on the website was categorized by the health care delivery setting to which its measures belonged (i.e., hospital care, long-term care, or home care), and by the types of measures (e.g., clinical, patient perception of care, health status). Since the website listed more than 250 ORYX vendors, the project team selected the relevant subset of vendors to contact.

Each vendor was categorized into one of three groups: 1) vendors with clearly relevant QIs whom the project team would target with letters, following up with second letters if no response arrived; 2) vendors with potentially relevant QIs whom the project team would initially contact but would not aggressively follow up; and 3) vendors with QIs that were irrelevant to this study whom the project team would not contact. After considering all the QI systems offered by these vendors, the project team found that any systems listed as applicable to long-term care settings automatically fell into category 1, as well as systems with names indicating relevance to post-acute care (for instance, "Post-Acute Support Systems"). Systems applicable to hospital care were assigned to category 2, if the types of measures they addressed included clinical or health status measures, using the reasoning that this hospital care category would capture post-acute care QIs. The project team eliminated systems that applied only to home care QIs or to hospital care QIs that did not address clinical or health status measures, and also dropped hospital care QIs whose name implied a population irrelevant to this study (e.g., pediatrics, perinatal care, and emergency room care). The selection process eliminated more than 50 systems from the initial list of over 250 ORYX vendors.

3.3.2 Initial and Follow-up Mailings to ORYX Vendors

Letters were sent to the 198 ORYX software vendors whose systems fell into categories 1 or 2, requesting information about the vendors' QI systems, and providing background on CMS's goal of developing a national quality monitoring system. The project team also provided a short form for each vendor to fill out, specifying the title of each QI, its description, numerator, denominator, MDS variables, and covariates. (See Appendix 2 for copies of letter and form.) The letter informed vendors that all QIs ultimately recommended for inclusion in CMS's QI system would belong in the public domain. Vendors who did not wish the project team to analyze their QIs had the option of returning an exclusion statement. The letter requested a response within 2½ weeks.

At the end of the waiting period, the project team mailed follow-up letters to vendors whose responses had not yet arrived, focusing on 127 vendors from category 1, those with clearly relevant long-term or post-acute care QIs.

3.3.3 Summary of Responses

Overall, 59 vendors responded to the mailing, of which 28 submitted the requested information about their QIs. Of these, 112 QIs from 15 vendors were forwarded for review, while QIs from the remaining 13 vendors were not. (Included in the 112 QIs were 29 from the Center for Health Systems Research and Analysis, University of Wisconsin-Madison, previously identified through the literature search.) The selected QIs applied either to long-term care, post-acute care, or special populations. In addition, the project team looked for QIs that could be validated, contained statements of numerator, denominator, and covariates, and were MDS-based.

Of the 13 ORYX vendors with QIs not forwarded for review, four used the CHSRA measures. QIs from three vendors, The Arcon Group, Inc. (addressing functional assessment in acute, sub-acute, long term, post-acute, and home care); the New York Association of Homes & Services for the Aging (defining and modifying a subset of the CHSRA QIs); and National Healthcare Corporation may be useful in the future for developing new QIs, but the project team excluded them from the initial selection process either because of a lack of information returned or a late arrival of information. Several of the other vendors used only mental health or hospital QIs, or focused on unrelated areas such as child health measurement systems, patient satisfaction measures, and neurological assessments. Others used non-MDS data or did not provide a numerator and denominator or a method of risk adjustment.

Finally, 16 vendors responded by stating that they had no QIs that would be useful to this study, and 15 vendors returned the exclusion statement, requesting that the project team not consider their QIs in this study. (See Appendix 2 for the complete list of ORYX vendors who responded to the mailings.)

3.3.4 Other Sources Contacted: Nurse Executives

The project team generated a list of prominent nurse executives involved in long-term care, using names of participants at a 1998 forum on long-term care issues in nursing, as well as project team members' personal contacts. Next, a letter was drafted to 22 of these nurse executives requesting information about their QI systems. Of the six who responded, all provided the requested QI information. Of the six responses, three used the CHSRA measures while three contained no QI information useful for this study. Table 3.1 below summarizes all of these efforts.

Table 3.1
Summary of QI Searches

Literature search	
Unduplicated hits	37
Reviewed	16
Internet search	
Unduplicated hits	4
ORYX vendors	
Contacted	198
Responded with information	28
Reviewed	15
Responded with no information	16
Requested the project team NOT use their QIs	15
Nurse executives	
Contacted	21
Responded with unduplicated information	6
Reviewed	1

An ORACLE database containing structured questions as well as open text boxes was used to describe the QIs. (See Appendix 3 for a copy of the ORACLE cataloguing form.) Reviewers classified each QI into one of four domains: functional, clinical complexity, psychosocial and pharmacotherapy. Information gathered included the use for each QI, the population using it, whether its basis was cross-sectional or longitudinal information, and the rationale for the definition. Reviewers also had to specify the technical aspects of the QI including the numerator, denominator, and risk adjustment method.

4.0 Selection of QIs for Empirical Analysis

QIs with the potential of being measured with MDS data were submitted for closer review by researchers and clinicians with expertise in the area. Reviewers followed a structured format for evaluating and documenting information on each QI. Overall, 143 QIs were reviewed and 31 were selected to undergo the empirical analyses. An additional 13 were forwarded for empirical testing because they were selected by CMS for use in the survey process. Below is a description of the review process and the criteria used to select QIs for empirical testing.

4.1 Reviews and Oracle Database

Review Process. The first step in the review process was the construction of review teams for each of the substantive domains into which the QIs had been classified (function, clinical complexity, psychosocial, pharmacotherapy). Each team consisted of clinicians with long-term care content expertise in the domain as well as experienced researchers aware of the technical issues associated with constructing QIs. Each team was also staffed with a coordinator who scheduled meetings via conference calls and who maintained minutes of deliberations and decisions.

All QIs that passed a cursory check for relevance and had a possibility of being operationally defined by MDS items were sent for review. Reviewers were project team members including clinicians with expertise in the specific content area the QI addressed. Reviewers were provided published articles or information from ORYX vendors and other sources. Reviewers received all articles and information pertinent to a specific area such as pressure ulcers, incontinence or functional decline. Articles often contained more than one QI and the same QI may have been described in multiple sources. Each different QI was entered as one record into a computerized tracking system — an ORACLE database. Reviewers made the determination of which QIs represented unique markers. In total 143 QIs were entered into the database. However, later clarifications from organizations revealed that certain records were duplicates. That is, other organizations submitted QIs that were in fact identical to CHSRA QIs.

The ORACLE database contained structured questions as well as open text boxes to describe the QI. (See Appendix 3 for a copy of the ORACLE cataloguing form.) Reviewers classified each QI into one of four domains: functional, clinical complexity, psychosocial and pharmacotherapy. Questions included how each QI was used, the population for which it was used, whether it was based on cross-sectional or longitudinal information and the rationale for the definition. Reviewers also had to specify the technical aspects of the QI including the numerator, denominator and risk adjustment method. The numerator refers to the upper portion of a fraction used to calculate a rate, proportion, or ratio i.e., patients who have the characteristic or outcome of interest. The denominator refers to the lower part of a fraction used to calculate the rate, proportion, or ratio. It may refer to those at risk of developing the outcome or characteristic of interest or may refer to all persons in a facility. Risk adjustment within the quality improvement literature may consist of one or more of three basic types: restricted denominator, separate calculation of QI rates within risk groups or strata, and use of multivariate adjustment modeling. In addition, the ORACLE database asked reviewers to document available information on the measurement properties of the QI, including reliability, validity and responsiveness. (Appendices 4 and 5 provide a detailed listing of the information catalogued on each QI, as well as the references used.)

Results of the review process. In some cases, it became clear during the selection process that there was insufficient information with which to initiate the analysis phase of this work. For instance, a QI developer might have published or provided the operational definition of the numerator and denominator of the QI, but not have provided the code for specific MDS items. In several cases, it was unclear how the developer handled missing values, and which version of the data collection instrument the developer utilized (MDS 1.0+ or 2.0, quarterly or full assessment, etc.). To ensure complete information about each QI, a table was constructed that summarized the missing items needed for analysis. Six developers received a letter via Federal Express requesting the missing information, and all six received follow-up calls.

There was a range of responses to these requests. One developer expressed interest in cooperating when contacted by phone, but never responded with the requested information. Another, it turned out, had passed away a few years ago, and the Steering Committee voted to drop his QIs from the analysis due to the inability to clarify his QIs. Other developers provided the information requested through a series of letters and calls; as more was learned about each QI and the information needed to model it using secondary data, the information exchange process was repeated. In most cases the project team was able to clarify any points of confusion. In the few cases where information was unavailable and a decision was made to pursue modeling of the QI, the Steering Committee agreed that project programmer/analysts would extrapolate the missing information to the best of their ability.

Overall, there was a paucity of information on the measurement properties of existing QIs. Nonetheless, there were many QIs which were in use in a variety of clinical settings or research projects, suggesting substantial content validity and perceived clinical utility.

Final selection criteria. The Steering Committee assumed responsibility for selecting which QIs were forwarded for empirical analysis. A protocol was determined whereby all decisions required at least a 4-2 majority vote. Based on the results of the review process, the Steering Committee determined that the minimum criteria for selection would be the presence of a clearly specified numerator and denominator, both of which could be operationally defined using MDS items. A priori the Steering Committee decided to give preference to QIs with some form of risk adjustment in order to permit a fairer comparison between facilities with different patient populations (or casemix). However, no QIs failed to be forwarded for empirical analysis solely on the basis of no risk adjustment. Consideration was also given to the perceived clinical relevance, presence of literature either supporting or opposing the concept and its relationship to quality, and whether the expected prevalence or incidence would be sufficient to function as a reliable QI over time.

For this task, QIs that rely upon data other than the Minimum Data Set (MDS) assessment form, such as CMS's Online Survey Certification and Reporting (OSCAR) assessment form, were not given as much weight as those based upon MDS data, for several reasons. First, reliance upon OSCAR data generated as part of state surveys is problematic because such data generate measures of structural and process quality with no resident-level risk adjustment possible for the few possible "outcome" measures. Secondly, QI systems relying upon record-based clinical data or resident surveys would require the introduction of new, large scale data collection efforts which would not have been tested or validated for some time. Since this project task emphasizes testing *existing* QIs to determine which are ready for generalized use, and they can be categorized as ready to use only if their underlying data are available, only those QIs based upon MDS-like, resident-level assessment data were considered.

Project staff and expert clinicians assisted in determining which QIs met the criteria, but the final decision was made by a vote of the Steering Committee. The main reason for not submitting QIs for analysis was lack of clear definition of numerator and denominators, and the inability to define the QIs based on MDS data. Table 4.1 presents all of the QIs that meet the criteria, classified within four domains: functional, clinical complexity, psychosocial and pharmacotherapy. Under each domain, a series of concepts are enumerated, and under these the "brand name" of the selected QIs are presented. Most QIs that met study criteria and were forwarded for analysis were derived either from the Center for Health Service Research and Analysis at the University of Wisconsin or from LTCQ's Q-Metrics© Information Advisory System. Both QI systems have been used by facilities in multiple states for many years and are certified as ORYX vendors under the JCAHO outcomes measurement initiative. The CHSRA QIs have been adopted for use by CMS in monitoring LTC quality.

Though the literature search did field some existing indicators of quality of life and satisfaction, and organization processes, more were found to meet the study criteria. No post-acute QIs were proposed for validation because of the limitations of our data. Most of the post-acute QIs required admission and discharge assessments which were not available in the current databases. In addition, few post-acute patients were represented in the database. The development and validation of post-acute QIs will be addressed in later project tasks.

Finally, although some did not meet the study criteria for proceeding to analysis, all 24 CHSRA QIs incorporated into CMS's new survey process were forwarded for empirical testing.

Table 4.1
QIs Selected for Further Analysis, by Domain

Domain	QI Name	Developers
Functional	ADL decline	CHSRA LTCQ Mukamel
	Bedfast	CHSRA*
	Cognition	CHSRA LTCQ
	Communication	LTCQ
	Decline in ROM	CHSRA HCDS
	Locomotion	LTCQ
Clinical Complexity	Dehydration	CHSRA*
	Falls	CHSRA LTCQ
	Fecal impaction	CHSRA*
	Incontinence	CHSRA (bladder or bowel stratified) CHSRA (bladder or bowel without toileting plan)* LTCQ (bladder and bowel)
	Indwelling urinary catheters	CHSRA (high/low risk) LTCQ
	New fracture	CHSRA*
	Pain	LTCQ
	Pressure ulcers	LTCQ CHSRA Mukamel
	Restraints	CHSRA* LTCQ Mukamel
	Tube feeding	CHSRA* Ramsey
	Urinary tract infection	CHSRA*
	Weight loss	CHSRA* LTCQ
Psychosocial	Behavior	CHSRA LTCQ
	Depression with no treatment	CHSRA
	Little or no activity	CHSRA*
	Mood	CHSRA* LTCQ
	Personal relationships	LTCQ
Pharmacotherapy	Hypnotic use >2	CHSRA*
	Medication number	CHSRA
	Prevalence of anti-psychotics	CHSRA
	Prevalence of anti-anxiety/hypnotic use	CHSRA

*CHSRA QIs not subject to the selection process (i.e., were not required to meet minimum criteria for further empirical analyses), but forwarded for analysis due to their current use in the long-term care survey process.

5.0 Methods Used in Initial Empirical Analyses

To examine aspects of the empirical and statistical performance of the QIs, the project team subjected all selected QIs to an extensive set of analyses. The sections below contain descriptions of each of the analysis steps. Chapter 6 contains the results of these empirical analyses, presented as a series of tables that describe the performance of each candidate QI with respect to: prevalence (at the resident level as well as the aggregated facility level); adequacy of risk adjustment; and stability of the QIs over time.

5.1 Description of Databases

Quarterly and annual MDS files were available for Kansas, Maine, New York, South Dakota, and Vermont. These data files contain all variables appearing on each type of MDS assessment, including full admission, annual assessment, and somewhat abbreviated quarterly assessments. Using these data, the project team created the following files:

- Quarterly Resident Files: These resident files contain four quarters of CY1996 data. There are from one to four assessments for residents who had at least one assessment over the four-quarter period. The most recent assessment identified may be either a quarterly (Q) or a full (F). Each quarter contains one record for each resident assessed at the facility during that quarter. If the resident had more than one assessment during a quarter, the priority for selecting which assessment to include in the file was to first include the full assessment and second, include the most recent quarterly assessment during the quarter;

- Annual Resident Files: These resident files contain two full assessments: the most recent full assessment for individuals with a full assessment during the last six months of 1996; and the earliest prior full assessment at least 12 months but not more than 18 months prior. Thus, if a resident was in the facility during the prior period, but not in the facility during the last six months of 1996, that case was not eligible for this file. In addition, a resident's initial assessment during the last six months of 1996 led to that resident's exclusion because that resident had no earlier assessment; and

- Quarterly and Annual Facility Aggregate Counterparts to the Resident Files: These files contain aggregated measures made at the facility level, based on the residents in the quarterly and annual files. A minimum of 20 aggregated cases for the QI under review were required in order for the facility to be included in the analysis of that QI.

These files contain data collected using different versions of the MDS: MDS 1.0, 2.0, and MDS PLUS. There are two variants of the PLUS version of the MDS: one labeled as + (a version of MDS 1.0, dated 12/90, created for use in the Nursing Home Casemix and Quality demonstration) and the second labeled ++, an update of +, dated 12/92. Table 5.1 describes the version of data available in the quarterly file for each state. Note that wherever possible the project team collapsed the items from the different versions into a common variable set. Research necessity sometimes required distinct versions of the variables.

The rationale for conducting empirical "validation" analyses of the selected QIs in multiple states relates to the fact that experience working with these data has shown that there are substantial inter-state differences in the prevalence of certain clinical conditions or outcomes such as functional change rates. This had alerted the team to the real possibility that data from different states might yield different answers to the question of the adequacy of the QI. While experience with these data suggested that some of the observed inter-state differences were attributable to real differences in the populations of nursing home residents from state to state, the project team recognized that there were also substantial inter-state differences in the measurement and assessment approaches used by nursing facility staff. Thus, we expected to observe **both** inter- and intra-state differences in the approach to measurement of clinically relevant phenomenon affecting QI performance.

Table 5.1
Versions of MDS Data Available in the Quarterly Assessment Data File for Each State

State	Jan-Mar (a)	Apr-Jun (b)	Jul-Sep (c)	Oct-Dec (d)
Vermont	+	+	2.0	2.0
Kansas	++	++	++	++
Maine	+	+	+	+
New York	+	+	+	+
South Dakota	+	+	+	+

Creating Cross-sectional and Longitudinal QIs

For prevalence QIs, the project team created the numerator and denominator for each individual using the assessment in the current quarter (e.g., for pressure ulcers, the data indicating presence of a pressure ulcer was available in the quarter). For change scores, each resident included in the analysis had to have a legitimate assessment in both the current quarter and the previous quarter (e.g., a new incidence of pressure ulcer in the current quarter that was not present in the previous quarter). Covariates, to the extent they were included in the QI, were usually derived from the prior assessment. However, some variables used as covariates were available only when the assessment was a full assessment. For example, medical diagnosis as a covariate was available only on the prior full assessment. If the "baseline" quarter (i.e., the previous assessment) was not a full assessment, the research team used the diagnosis from the prior full assessment and these data were attributed to the baseline assessment.

5.2 Analysis Plan

5.2.1 Overview

For each QI tested, an analysis report summarizing the prevalence of the QI and its distribution within and across at least three states' population-based MDS data was generated. These reports may be found in Appendix 6. The project team examined facility QI rates within states; specifically, the mean rate and standard deviation, as well as the rates at the

10th, 50th, 80th, and 90th percentiles within each state. QIs with mean rates below two percent were "flagged" for further discussion as to whether they really represented sentinel events rather than a QI. The definition of a sentinel event was: a rare but critical phenomenon that might signify immediate danger of harm for residents.

The project team also examined the distribution of age, gender, race, and any risk factors specified within the definition of the QI. To identify variation among states, particularly for the relationship between resident age and the QI, the project team examined the relationship between the demographic characteristics of the residents and the prevalence of the QI. For those QIs with specified covariate, or stratification, structures, the project team examined the bivariate relationship between the covariates and the unadjusted QI prior to multivariate modeling or computation of rates within strata. Furthermore, prior to multivariate modeling, the project team examined the interrelationships among covariates for evidence of collinearity. The consistency of these relationships across states was then examined.

The project team entered all risk factors specified by the developers of the QIs into the multivariate models, then examined odds ratios to determine the direction of the effect on the observed outcome, statistical significance,[1] and consistency across states. The latter was particularly important in that to be acceptable, a covariate had to show consistency in at least two state data sets.

The project team compared the aggregated facility-level adjusted rate of each QI with the observed (i.e., unadjusted) rate for that facility. For QIs which were stratified, the project team examined the correlation among facilities' ranking within state between their high and low risk QI score. All facility-level QI rates were examined for stability over time (across quarters) in two ways: 1) by correlating rank order of the deciles of the two sets of QI scores; and 2) by comparing movement on terciles of the QI distribution (e.g., whether facilities in the top one-third of the state distribution in one quarter retained that designation in the next quarter).

By hypothesis, facility aggregates of organizational factors such as staffing ratios and other aggregate measures of facility type were markers of better quality, and were thus used for secondary validation of the existing QIs. The correlations between the QIs and the aggregate variables were, however, of relatively low magnitude and inconsistent across states. The project team had not expected strong relationships due to the lack of objective, comprehensive measures of quality. Nonetheless, the lack of association was disappointing, requiring abandonment of this particular validation strategy. The project team will conduct validation of some existing and all newly-developed QIs under this project, during the subsequent primary data collection phase.

5.2.2 Methodology for Constructing Risk-adjusted QIs

As noted above, most of the existing QIs identified by the research team have some form of casemix risk adjustment using either stratification or some form of regression adjustment. The former simply classifies residents into one or more groups (strata) based on one, or a

[1] Statistical significance was defined as odds ratios with a 95 percent confidence interval that did not include one.

composite, of individual clinical factors, e.g., diagnosis, mobility. While multiple strata are possible, the more strata in a QI, the smaller the number of cases included in the measure, and therefore the less statistically stable and reliable the quality estimate for a facility. This especially becomes a problem when more than one dimension is necessary to stratify residents, as the number of groups is the product of the dimensions for each stratifying variable. Nonetheless, risk strata have the advantage of being readily understandable.

In some instances the number of risk strata needed to adequately adjust for known variation in residents' risk of a condition exceeds the precision those strata can reasonably provide. In such instances, multi-variable adjustment using some form of regression adjustment is necessary. All LTCQ QIs and several other QIs that were tested used this form of regression adjustment.

To calculate QI rates adjusted for resident-level risk factors, the project team used the method described by Berlowitz and his colleagues (1996). This technique only applies to those QIs relying on regression adjustment techniques. Using logistic regression, each resident's logarithmic odds of experiencing an event is modeled as a linear function of his or her clinical characteristics. After arithmetic transformation, a predicted event probability for each resident is calculated from this estimation. Those predicted probabilities are added up by facility and then divided by the number of residents at risk for the respective event to retrieve an expected event rate for the facility. The ratio of the actually observed event rate and the expected event rate is multiplied by the grand mean event rate, i.e., the event rate across all facilities, to give the risk-adjusted QI rate. The corresponding formula is:

$$QI_{adj} = (QI_{obs} / QI_{pred}) \times \text{grand mean}$$

Adjusted event rates based on this technique have the following useful properties:

1. The better the facility is doing compared to the model's prediction, the better (lower) is its QI rate;
2. The worse the facility is doing compared to the model's prediction, the worse (higher) is its QI rate;
3. If the facility's observed QI rate is equal to 0, its adjusted QI rate is also equal to 0; and
4. The average adjusted QI rate is close to the average observed QI rate.

With this approach, for facilities with different observed rates of QIs, the project team could adjust for part of the difference attributable to the influence of known risk factors not under the control of the facility. For example, as mobility decreases with age, a facility's rate of immobile residents will increase with the mean age of its residents, all other things equal. Thus, a direct comparison of facilities without adjustment for age would penalize/reward some facilities for factors beyond their control.

The predicted probabilities based on the regression technique show what the rates of QIs in the facilities were with each facility serving a comparable mix of residents, at least taking the covariates in the model into account. If some facility has an observed rate higher than predicted by the model, it would mean that this facility is performing worse than it should, given its casemix (with respect to the QI under scrutiny). And vice versa, if the observed rate is lower than predicted, this facility is better than would be expected based on the distribution of residents in the home. For example, if the residents are older, mobility is expected to be somewhat worse, all other factors being equal. It would be unreasonable to penalize facilities for an unfavorable casemix.

5.2.3 Aggregating QIs and Estimating Standard Errors of Estimate

As noted earlier, there is a well-established relationship between the number of observations on which an estimate is based and its accuracy, or confidence level. Creating a QI for a given facility requires aggregating data about residents, events, or treatment processes to the level of the facility. Although all appropriate observations may be included in the aggregate measure being constructed, a sample still determines the measure, since observations may change over time for any number of different reasons. A facility QI in one month may differ from the facility QI on the same measurement concept in subsequent months, due either to real changes, or "sampling changes", or both. The sampling changes might be due to slight differences in the facility's pool of residents, or that different staff members are completing the MDS (this would actually be measurement error). Consequently, any single QI measure should only be considered as an indicator of possible quality problems rather than as an ipso facto measure of quality.

All estimates have a certain degree of associated error. In the case of constructing aggregated QIs, the best understanding of the cause of sampling error is the number of observations determining the QI. The larger the denominator determining the QI estimate, the more likely that the observed score is reasonably close to the "true" score. Table 5.2 below summarizes this relationship for a hypothetical QI that has a prevalence (or incidence) of only 5 percent.

Table 5.2
Relationship Between Sample Size and the Standard Error of Estimate for a Hypothetical QI with Incidence of 0.05

Number of Observations	Standard Error	95% CI (Binomial Exact)
10	.09	.002 - .444
20	.05	.001 - .245
30	.04	.008 - .223
50	.03	.01 - .16
100	.02	.01 - .11
200	.01	.02 - .09
500	.009	.03 - .07

Similarly, wide confidence intervals exist for proportions of .25 and .45. For a QI (e.g., cognitive impairment), with a prevalence of .45, the confidence intervals surrounding an estimate based on only 20 residents are .23 - .68. For this reason, comparative analyses undertaken as part of the empirical validation analyses of specific QIs required that facilities have a minimum of 20 legitimate observations for a specific QI.

5.2.4 Assessing the Stability of QIs over Successive Quarters

The problem of potentially unstable estimates of the prevalence or incidence of QIs, particularly in smaller facilities, required systematic examination of the stability of QI estimates over the course of successive measurement quarters. The results of these analyses were influential in respective recommendations about the appropriateness of one QI over another. As noted earlier, prevalence QIs frequently had most of the same patients in the numerators and denominators of measures taken in successive quarters. Such measures would obviously have apparently high stability precisely because residents continuing to reside in a given facility are unlikely to change greatly in the absence of an acute event. However, the concept of stability refers not just to the stability of individual residents' condition, but rather how the facility meets the changing needs of its residents. It is this latter concept that was given greatest weight in attributing meaning to the stability of QIs.

The project team used two different approaches to estimate the stability of a given QI. Within a given state, the first was the classification of the underlying facility-level QI distribution into deciles as well as terciles, which was important for two reasons. First, the distributions tended to be non-normal and skewed. Second, it was important to know whether more facilities improved or declined.

The second approach to estimating stability was that for each QI tested the Spearman rank order correlation between successive quarters was calculated. Separately by facility size, the project team also calculated the percentage of facilities that changed two as well as three deciles on each QI over a three-month period. Finally, a cross-tabular analysis of the QI terciles in one quarter with the QI terciles in the next was constructed. The project team conducted all these analyses separately within state, with each QI tested on at least two states, and most often on three states.

6.0 Results of Initial Empirical Analyses and Selection of Recommended QIs

6.1 Overview of Results of QI Testing

Each QI selected through the process described in Chapter 4 was subjected to a series of analysis steps as described in the preceding section. Following the statistical analyses the project Steering Committee (SC) discussed and thoroughly reviewed each QI in terms of its distributional characteristics, stability and cross-state consistency. The balancing perspective adopted in reviewing QIs was not to look for QIs which met all established performance criteria since none were really without problems. Rather, the Steering Committee considered those QIs that minimized the problems of ascertainment bias, censoring, skewed distribution, casemix adjustment and QIs that resulted in dropping too many facilities due to insufficient sample size. On the other hand, in certain instances the concepts identified in QIs that may not have met standards on statistical grounds were deemed to be so important from a clinical or operational perspective that the Steering Committee agreed to include them in the list of previously recommended QIs. The results are presented primarily in tabular form. Table 6.1 is a summary table of the number of QIs within each area submitted for testing and the number accepted and rejected.

The descriptive results and additional information on the accepted and rejected QIs are presented in Tables 6.2.1-4 and Tables 6.3.1.-4, respectively. The tables are organized by domain in the following sequence: 1. functional, 2. clinical complexity, 3. psycho-social and 4. pharmacotherapy. Reporting forms containing the details of the empirical analyses are provided in Appendix 6.

6.1.1 Summary of QIs Reviewed

Table 6.1 summarizes the QIs reviewed and accepted following empirical analyses. Within each domain, QIs were categorized into those that were "prevalence-based or cross-sectional" as opposed to those that were based upon an incident event or the rate of deterioration in residents' status. In light of the difficulty of establishing risk adjustment models for prevalence QIs, more of the "change in status" QIs were recommended than were prevalence measures. Overall, 26 out of the 44 evaluated QIs were recommended for use after the first series of empirical analyses.

Appendix 7 provides the definition and functional form of each reviewed QI, and Appendix 8 describes each QI recommended for use by CMS in layman's terms.

Table 6.1
Summary of QIs which Underwent Empirical Testing

QI Domain	CROSS-SECTIONAL QIs		CHANGE IN STATUS QIs	
	Total Reviewed	Number Initially Accepted	Total Reviewed	Number Initially Accepted
I. FUNCTIONAL STATUS				
Communication/Cognition				
Communication	0	0	1	1
Cognition	0	0	2	1
ADL Status				
Bedfast	1	0	0	0
Locomotion	0	0	1	1
ADL	0	0	3	1
Range of Motion	0	0	2	0
SUMMARY: FUNCTIONAL QIs	**1**	**0**	**9**	**4**
II. CLINICAL COMPLEXITY				
Continence Related				
Bladder/Bowel	2	1	2	2
Fecal Impaction	1	0	0	0
Catheter	1	1	1	1
UTI	1	1	0	0
Nutrition/Hydration				
Dehydrated	1	0	0	0
Weight Loss	1	0	1	1
Tube Feeding	2	1	0	0
Restraints	2	1	1	0
Falls/Fracture				
Falls	0	0	2	1
New Fractures	0	0	1	0
Pain	0	0	1	1
Pressure Ulcers	2	1	1	1
SUMMARY: CLINICAL QIs	**13**	**6**	**10**	**7**
III. PSYCHOSOCIAL				
Mood	2	2	1	1
Behavior	1	1	1	1
Activity	1	1	0	0
Personal Relationships	0	0	1	1
SUMMARY: PSYCHOSOCIAL QIs	**4**	**4**	**3**	**3**
IV. PHARMACOTHERAPY				
Anti-anxiety/hypnotics	2	1	0	0
Anti-psychotic	1	1	0	0
9 or more medications	1	0	0	0
SUMMARY: PHARMACOTHERAPY QIs	**4**	**2**	**0**	**0**
V. GRAND SUMMARY	**22**	**12**	**22**	**14**

6.2 QIs Accepted for Use by CMS

Tables 6.2.1-4 present information on the QIs recommended for use within each domain. The majority of these results are presented in tabular form.

These tables show the descriptive statistics for each QI together with the main reasons for accepting the QI. Specifically, the tables present:

(1) The name of the QI and original developer;

(2) The average facility rates for each state. This raw or unadjusted rate refers to the rate or proportion defined by the numerator and denominator alone;

(3) If the QI was risk-adjusted, the adjusted facility averages are displayed. For QIs stratified by high and low risk groups, the stratum-specific rates are presented separately. For QIs that include a regression-based risk adjustment (such as the LTCQ QIs), adjusted rates were computed using the methods described in Chapter 5. Briefly, the adjusted rate represents a ratio of the observed incidence rate to the expected rate multiplied by the state average rate. The expected rate is based on the average predicted probability of incidence within the facility, given the type of residents residing in the facility. Facilities with an observed rate that is lower than predicted by the model would be presumed to be performing better than expected based on the casemix within that facility. As expected by the nature of the adjustment method, average raw and adjusted QI rates are almost identical;

(4) The facility rates at the 10th and 90th percentiles for each state;

(5) The Spearman rank correlation coefficients between decile rankings over time within each state used to evaluate QI stability;

(6) The risk adjustment method used for the respective QI; and

(7) The primary rationale for accepting or rejecting the QI.

6.2.1 Functional QIs Accepted for Use

Four functional QIs were recommended for use in the areas of functional decline, cognition, communication and locomotion. Each QI represented a decline in status from one assessment to the next quarterly assessment and excluded those who could not decline further because they had maximal scores at the start of the time period. **Functional decline** was measured by a two-level decline in eating, bed mobility, transfer and toileting or a one-level decline in two or more of these "late-loss" activities of daily living (ADL). **Decline in cognitive function** was measured by any deterioration in the Cognitive Performance Scale (CPS) over time. Similarly, **deterioration in communication** was based on a decline in the MDS communication scale. **Locomotion** was measured by combining wheelchair mobility and walking into a single variable and examining residents with some degree of independence who became more dependent within the next 90 days.

Table 6.2.1 summarizes the results of the empirical analyses for these four accepted functional QIs in terms of the rates, consistency across states and across time periods, type of risk adjustment, and the primary reason for recommending the QI for use.

None of the four QIs had a low rate of occurrence, as all were .08 or above. Mean adjusted and unadjusted rates for the QIs were very similar. There was marked variation at the 90th percentile rate across states, with a doubling of the rate in some states.

Only an unadjusted rate is shown for the ADL decline as it has no risk adjustment beyond restricting the resident pool to residents with some degree of independent function at the baseline assessment. Cognition, communication and locomotion had regression-based risk adjustment models with 16, 10 and 12 covariates, respectively. Column 6 presents the Spearman rank order correlation between decile rankings of facilities within state over successive quarters. The percentage of facilities that changed three deciles on the QI over a three-month period was also calculated separately by facility size. Analyses were conducted separately within state, with each QI tested on at least two states and most often on two or three states. The four accepted functional QIs show relatively low associations between scores over time, suggesting instability in rankings from quarter to quarter. Indeed, between 40-50 percent of facilities showed more than three decile changes over time in locomotion.

Lastly, Column 8 reviews the intended purpose of each QI and the primary reason for recommending the QI for use.

Table 6.2.1
Functional QIs Accepted for Further Validation.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
ADL decline (CHSRA)	VT KS NY	0.15 0.12 0.09		0.06 0.04 0.04	0.29 0.21 0.15	-0.01* 0.35* 0.41*	NONE	This QI measures functional decline by a two-point change in eating, bed mobility, transfer and toileting or a one-level decline in 2 or more late-loss activities of daily living (ADL). (Each of the MDS ADL variables is measured on a 5-point scale from 0-4). This QI was selected over the LTCQ QI because it was more conservative (i.e., required 2 points vs 1 point) in measuring a decline in function, suggesting less tendency for measurement error. One drawback of this QI is the lack of risk adjustment beyond exclusion of residents with complete dependency.
Cognition (LTCQ)	VT KS NY	0.16 0.10 0.09	0.17 0.10 0.09	0.06 0.03 0.04	0.32 0.19 0.14	0.39* 0.41* 0.33*	Regression Based Adjustment (16 Covariates)	This QI measures worsening of cognitive function by monitoring any deterioration in cognitive performance scale (CPS) scores. The LTCQ QI was chosen ahead of the CHSRA QI because it was more broadly applicable to nursing facility residents. The CHSRA QI measures the incidence of cognitive decline and thus excludes the majority of nursing facility residents i.e., those with some degree of cognitive decline in the previous assessment.
Communication (LTCQ)	VT KS NY	0.15 0.09 0.08	0.15 0.09 0.08	0.04 0.01 0.03	0.31 0.17 0.14	0.21* 0.43* 0.41*	Regression Based Adjustment (10 Covariates)	This QI measures a decline in communication performance based on the MDS functional communication scale (making self-understood and understanding others). This QI was recommended because it was the only one to address communication, an area deemed important, and because the risk adjustment model contained factors consistent with the literature.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

Table 6.2.1
Functional QIs Accepted for Further Validation.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decisi on
		Mean	Mean	10th %	90th %			
Locomotion (LTCQ)	ME KS NY	0.23 0.12 0.11	0.23 0.12 0.11	0.09 0.03 0.04	0.42 0.22 0.19	0.38** 0.49** 0.47**	Regression Based Adjustment (12 Covariates)	This QI looks at the decline in resident's performance in mobility, either self-propelled in a wheelchair or when walking. The QI was recommended for use as it was the only one to address prevention of greater dependency in locomotion.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

6.2.2 Clinical Complexity QIs Accepted for Use

A total of twenty-three clinical QIs were empirically analyzed, 13 cross-sectional representing the prevalence of a clinical condition, and 10 longitudinal representing change in status (i.e., incidence (new) or worsening of a condition). The Steering Committee initially voted to accept a total of 13 QIs (6 cross-sectional; 7 change in status) for recommended use by CMS. The accepted clinical QIs, which can be conceptually characterized as either representing resident symptoms/clinical conditions or clinical processes of care, are further described below. The results of the empirical analyses are summarized in Table 6.2.2.

QIs related to resident symptoms or clinical conditions

Nine accepted measures represented QIs related specifically to resident symptoms or clinical conditions (i.e., incontinence (3), urinary tract infection (1), pressure ulcer (2), falls (1), pain (1), and weight loss (1)). Three measures of incontinence were accepted, one cross-sectional and two change of status measures. The **Bowel and Bladder (CHSRA)** QI measures the prevalence of frequent or greater incontinence of either type. It is simple to follow, has face validity, and is able to distinguish residents at high or low risk for the condition. The **Bladder Incontinence (LTCQ)** QI measures the incidence or worsening of bladder incontinence, a common sign of potentially reversible or treatable conditions in the nursing facility population (e.g., delirium, urinary tract infection, joint pain limiting self-toileting ability). The **Bowel Incontinence (LTCQ)** QI measures the incidence or worsening of bowel incontinence, a potentially treatable problem that may be associated with underlying constipation, fecal impaction, or laxative use.

Table 6.2.2 shows that new or worsening incontinence is relatively common across states. While some decline in bladder or bowel continence may not be reversible or manageable in the latter stages of disease (e.g., dementia, terminal illness) these QIs as operationalized appear to have the capacity to identify facilities where there may be a quality problem. **Urinary Tract Infection (UTI)** is a common problem among frail, debilitated nursing facility residents, often leading to hospitalization and poor quality outcomes (e.g., delirium, incontinence, falls). Though the project team recommends exploration of risk adjustment, this CHSRA QI addresses an important dimension of nursing home quality. There are some problems with this QI; for example, although accurately defined in the RAI User's Manual, the MDS coding convention requires that the infection (item) be "symptomatic...and have current supporting documentation and significant laboratory findings in the medical record". Therefore, the possibility of measurement error by clinicians (or underreporting) is great. The item may be "gameable" by facilities (i.e., if the condition is not readily observable, it may not be recorded by staff) and may, therefore, be difficult for long-term care surveyors to monitor. Because adequate treatment often precipitates hospitalization, the risk of censoring bias is high. Finally, this QI lacks risk adjustment. This QI yields stable overall mean rates across states, but there is wide variation across states in the percentage of facilities with no urinary tract infections.

Two measures for **pressure ulcer** were accepted, one cross-sectional **(CHSRA)** and the other a change in status measure **(LTCQ)**. The CHSRA QI utilizes a high risk/low risk adjustment procedure to identify residents with any stage pressure ulcer on the most recent assessment. As it excludes new admissions and readmissions it focuses attention on the prevalence of ulcers occurring in the facility. The LTCQ QI that measures incidence or worsening of a pressure ulcer uses a risk-adjusted model with 10 covariates.

LTCQ QIs for **new or worsening pain, falls,** and **weight loss** are all QIs that represent common symptoms of potentially treatable or manageable underlying conditions, or in the case of pain and weight loss, conditions that would prompt palliative measures towards the end of life. Although there were some concerns with the specification of the pain and weight loss QIs, the Steering Committee determined that because they were both serious quality issues in nursing facilities, and they were the only covariate-adjusted QIs representing these concepts, they should be adopted for use by CMS at this time because they serve to identify facilities that have problems in these areas.

QIs Related to Clinical Processes of Care

Four of the accepted QIs represented clinical care processes: physical restraints (1), feeding tubes (1) and indwelling catheters (2).

Physical restraint use (CHSRA) measures the prevalence of daily use of limb or trunk restraint or chair that prevents rising. There is variation in the facility rates of daily restraint use across states and the rate at the 90th percentile varied from 0.13-0.23 across states. Although this QI is not risk- adjusted it does identify facilities with higher than average use.

The **feeding tube QI (Ramsey)** measures the prevalence of feeding tube use. The risk adjustment model includes 10 covariates to account primarily for persons with neurological and physical disorders at risk of the outcome. Even with risk adjustment there is wide variation in the overall rates of feeding tube use.

Two measures of **Indwelling Catheter (CHSRA and LTCQ)** were analyzed. The unadjusted CHSRA QI measures the prevalence of catheter use whereas the covariate-adjusted LTCQ QI measures the incidence of new catheters since the prior assessment. Both yield variation in distribution rates across facilities. Because indwelling catheters are associated with iatrogenesis and morbid outcomes in this population and these are important QIs over which facilities have some control, the Steering Committee recommends both QIs for use at this time.

Table 6.2.2
Clinical QIs Accepted for Further Validation.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Prevalence of bladder and bowel incontinence (CHSRA)	VT	0.60	*High Risk:* 0.93	0.85	1.00	0.55	High/Low Risk Groups (2 Covariates) Restricted Resident Pool	This prevalence QI for frequent or greater incontinence is simple to follow and performs well across states. The covariates have high Odds Ratios but the model is underspecified in view of very high rates in the low risk group (i.e., concern for false negatives). The QI rate is stable across time and there is a low-moderate correlation between ranking of low risk group and high risk group within facilities (.35-.45). There is evidence of some interstate variation. Finally for the high-risk group, 14% to 39% of facilities cannot be scored because of lack of 20 or more residents of this type in thefacility.
			Low Risk: 0.45	0.31	0.60	0.81		
	ME	0.65	*High Risk:* 0.91	0.80	1.00	0.73		
			Low Risk: 0.47	0.30	0.64	0.73		
	KS	0.45	*High Risk:* 0.86	0.70	1.00	0.78		
			Low Risk: 0.31	0.14	0.47	0.85		
	NY	0.60	*High Risk:* 0.95	0.89	1.00	0.74		
			Low Risk: 0.47	0.29	0.66	0.90		
	SD	0.46	*High Risk:* 0.88	0.71	1.00	0.78		
			Low Risk: 0.34	0.21	0.48	0.82		
Bladder Incontinence (LTCQ)	VT	0.20	0.20	0.08	0.30	0.43	Regression Based Adjustment (18 Covariates) Restricted Resident Pool	This QI measures the incidence or worsening of bladder incontinence, a common, though treatable problem in the nursing facility population. Analyses show good distribution of the QI across states. The covariate model has face validity and the estimated coefficients on the covariates have the expected sign. Although the correlations between the raw and adjusted scores are high, the adjustment scatterplots show that some facilities did change their scores and rankings following the adjustment procedure. Because of only moderate inter-time period QI correlations, need to explore pooling QI measures from multiple time periods in creating bench-marked QI estimates for facilities. Finally, because some facilities will have many residents who are fully incontinent, about 10% of facilities cannot be scored because there are fewer than 20 residents who are at risk of any decline.
	ME	0.20	0.20	0.08	0.31	0.30		
	KS	0.14	0.14	0.05	0.24	0.34		
	NY	0.14	0.14	0.07	0.22	0.33		
	SD	0.15	0.15	0.06	0.25	0.34		
Bowel Incontinence (LTCQ)	VT	0.17	0.17	0.07	0.28	0.03	Regression Based Adjustment (15 Covariates) Restricted Resident Pool	This QI measures the incidence or worsening of bowel incontinence, a treatable problem among many nursing facility residents. The covariate model has face validity and the signs on coefficients for all but two covariates are in the expected direction. Although the correlations between the raw and adjusted scores are high, the adjustment scatterplots show that some facilities did change their scores and rankings following the adjustment procedure.
	ME	0.19	0.19	0.04	0.31	0.51		
	KS	0.10	0.11	0.02	0.20	0.38		
	SD	0.13	0.13	0.04	0.23	0.36		
	NY	0.12	0.12	0.05	0.19	0.37		

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

Table 6.2.2
Clinical QIs Accepted for Further Validation.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Urinary Tract Infection (CHSRA)	VT	0.07		0.02	0.14	0.56	NONE	Urinary tract infection (UTI) is a common problem among frail, debilitated nursing facility residents, often leading to hospitalization and poor quality outcomes (e.g., delirium, incontinence, falls). The incidence of UTI is also strongly associated with use of chronic indwelling catheters.
	ME	0.07		0.00	0.13	0.50		
	KS	0.07		0.00	0.15	0.63		
	NY	0.07		0.02	0.12	0.68		
	SD	0.07		0.02	0.13	0.49		
Prevalence of Stage 1-4 Pressure Ulcers (CHSRA)	VT	0.09	*High Risk:* 0.12	0.03	0.22		Restricted Resident Pool (4 Covariates)	This adjusted point prevalence QI is easy to understand. The rates are considered separately for residents at high and low risk for developing pressure ulcers. There is moderate correlation between the low and high risk QI scores for the same facilities (0.31-0.63), meaning that if a facility performs well with those at low risk, there is some tendency for the facility to do well with those at high risk.
			Low Risk: 0.04	0.08	0.08			
	KS	0.07	*High Risk:* 0.13	0.00	0.24			
			Low Risk: 0.02	0.07	0.07			
	NY	0.10	*High Risk:* 0.16	0.07	0.25			
			Low Risk: 0.03	0.08	0.08			
Pressure Ulcers (LTCQ)	VT	0.05	0.06	0.00	0.14	0.96	Regression Based Adjustment (10 Covariates)	This covariate-adjusted QI measures pressure ulcer incidence or worsening. This QI is moderately stable over time; findings show that facilities with low rates are likely to continue to have low rates over the next quarter. Although the mean QI rates and adjustment findings are comparable to the Mukamel Pressure Ulcer QI, the covariate model is easier to follow and is more inclusive of relevant findings in recent literature.
	KS	0.04	0.05	0.00	0.10	0.97		
	NY	0.05	0.06	0.02	0.11	0.94		
Pain (LTCQ)	VT	0.08	0.08	0.00	0.19	—	Regression Based Adjustment (8 Covariates)	This QI measures the incidence/worsening of pain symptoms from baseline. There is wide variation in the overall pain rates across states and across adjusted percentile rankings of facilities within and across states. For the 10% of facilities with lowest rate, the incidence of pain over 90 days is basically zero. At the 90th percentile representing homes with the highest rates, the pain rates ranges from 0.07 to 0.27. Although the covariate model has face validity, it may be underspecified. The instability of the QI across time periods is of concern and there is a need to explore pooling QI measures from multiple time periods in creating benchmarked QI estimates for facilities. Despite these concerns, because pain is a serious quality issue in nursing facilities, and this QI was the only one available for our analysis, it was recommend for use at this time.
	NE	0.10	0.10	0.00	0.21	0.33		
	MS	0.07	0.07	0.00	0.27	—		
	TX	0.06		0.00	0.21	—		
	NY	0.03	0.03	0.00	0.07	0.29		

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

Table 6.2.2
Clinical QIs Accepted for Further Validation.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Falls (LTCQ)	VT KS NY	0.17 0.15 0.11	0.23 0.19 0.15	0.14 0.09 0.09	0.34 0.29 0.21	0.34 0.99 0.43	Regression Based Adjustment (17 Covariates)	Falls are common among the nursing facility population, particularly for those who have some independence in mobility. This covariate-adjusted QI, which measures the incidence of new falls in the last 30 days, is somewhat complicated in that the covariates reflect the "opportunity costs" of falling (i.e., residents with greater independence are at greater risk of falling than those who require help from others to complete tasks). The model adjusts for the situation in which facilities that have more independent residents in their casemix will have higher rates of falls.
Weight loss (LTCQ)	VT KS NY	0.08 0.09 0.07	0.08 0.08 0.07	0.02 0.02 0.00	0.15 0.15 0.14	0.21 0.26 0.20	Regression Based Adjustment (6 Covariates)	This covariate-adjusted QI measures incidence of new weight loss between quarters. The adjusted overall mean rates are stable across states, as are the rates at the 90th percentile. The interperiod correlations are very low, indicating that weight loss is a random event. Because weight loss is an important marker of quality problems in nursing facilities, and this QI is a risk- adjusted measure, the SC recommends it for use by HCFA.
Prevalence of daily physical restraints (CHSRA)	VT KS NY	0.08 0.05 0.09		0.00 0.00 0.00	0.21 0.13 0.23	0.87 0.83 0.94	NONE	There is variation in the rate of daily restraint use across states and across percentile rankings of facilities within and across states (for the 10% of facilities with the lowest use rates, no residents are in restraints; for 90th percentile representing facilities with the highest use, the proportion ranges from 0.13 -0.23). This QI is not risk adjusted; the SC believed that in view of the physical and psychosocial hazards associated with the use of physical restraints, survey teams should address the issue of restraints, and facilities with higher rates warrant particular scrutiny.
Prevalence of feeding tubes (Ramsey)	VT KS NY	0.03 0.02 0.08	0.04 0.03 0.08	0.00 0.00 0.00	0.10 0.08 0.17	0.65 0.59 0.73	Regression Based Adjustment (10 Covariates)	There is wide variation in the overall rates of feeding tube use across states. For the 10% of facilities with lowest use rates, no residents use feeding tubes. At the 90th percentile of facilities representing homes with the highest rates, residents use of feeding tubes ranges from 0.08 to 0.17. This prevalence QI is risk-adjusted and the model has face validity.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

Table 6.2.2
Clinical QIs Accepted for Further Validation.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Prevalence of indwelling catheter (CHSRA)	VT	0.05		0.00	0.09	0.83	NONE	This unadjusted, point prevalence QI yields low overall mean facility use rates across states. For the 10% of facilities with the lowest use rates, in each state almost all facilities have a zero use rate. At the other extreme, for the 10% of facilities with the highest use rates, between 0.09 and 0.11 of residents have an indwelling catheter. It is not clear to what extent casemix differences may explain the higher rates, because the QI is not casemix-adjusted. Because indwelling catheters are associated with iatrogenesis and morbid outcomes in this population, this is an important QI over which facilities have some control and facilities with higher rates will warrant particular scrutiny.
	ME	0.06		0.00	0.11	0.80		
	KS	0.05		0.00	0.10	0.72		
	NY	0.05		0.01	0.09	0.80		
	SD	0.05		0.00	0.11	0.83		
Indwelling catheter (LTCQ)	VT	0.01	0.00	0.00	0.01	[1] (see below)	Regression Based Adjustment (11 Covariates)	This covariate-adjusted QI measures the incidence of (new) indwelling catheters. The overall mean raw and adjusted rates are very low and there are many facilities with zero incidence. Because indwelling catheters are associated with iatrogenesis and morbid outcomes in this population, this is an important QI over which facilities have some control. The SC recommends using this QI as a sentinel event.
	ME	0.01	0.01	0.00	0.03			
	KS	0.01	0.01	0.00	0.03			
	NY	0.01	0.01	0.00	0.03			
	SD	0.01	0.01	0.00	0.03			

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

[1] Cannot perform stability analysis because of a lack of sufficient time periods of data (annual rather than quarterly).

6.2.3 Psychosocial QIs Accepted for Further Validation

Seven psychosocial QIs were recommended for use, four prevalence and three change QIs: Behavior (CHSRA), Behavior (LTCQ), Mood (CHSRA), Mood (LTCQ), Mood with no treatment (CHSRA), Little or No Activity (CHSRA), and Personal Relationships (LTCQ). Table 6.2.3 displays the results of the empirical analyses of these QIs.

The **Behavior (CHSRA)** prevalence QI utilizes a high risk/low risk adjustment procedure based on only 3 covariates. Nevertheless, the substantial differences of rates between the high and the low risk groups suggest that the covariates effectively stratify residents. Future work on this QI will focus on potential risk adjustment covariates beyond the high/low dichotomy.

The **Behavior (LTCQ)** change in status QI uses a covariate model and initially includes 21 covariates in the adjustment model. After reviewing the analyses, the Steering Committee recommended that the QI be brought forward for further validation work but with fewer covariates in the model. It was recommended that seven covariates be dropped. Some of these MDS items are considered service variables (i.e., drug use and restraint use) and thus not advisable for national implementation. The Steering Committee believed that, with these refinements to the covariate model, the LTCQ Behavior QI would be a strong QI of decline in behavior for immediate use by CMS.

The cross-sectional **Mood (CHSRA)** QI is not risk-adjusted. Despite its lack of risk adjustment, the Steering Committee recommended that it be adopted for use by CMS, as the QI has utility in detecting symptoms of depression among nursing facility residents. Future work on this QI will focus on potential risk adjustment covariates.

The **Mood (LTCQ)** change in status QI uses a 17-item covariate model. The Steering Committee recommended that the QI be brought forward but with fewer covariates in the model. It was recommended that seven covariates be dropped: antipsychotic drug use; antidepressant drug use; interest in reading/writing, leaves 25 percent of food uneaten; hemiplegia; loss of friend/family member; and internal bleeding. The Steering Committee believed that, with these refinements, this would be a good QI of decline in mood for immediate use by CMS.

The **Mood with no treatment (CHSRA)** QI is not risk-adjusted. Despite its lack of risk adjustment, it empirically demonstrates utility in detecting untreated symptoms of depression among nursing facility residents. This QI may be prone to ascertainment bias which means that some facilities will have higher rates because of better assessment. Ascertainment bias may partly explain the wide distribution of the facility rates within and across states. Future work on this QI will focus on potential risk adjustment covariates.

The **Personal Relationships (LTCQ)** QI is based on the annual MDS assessment. This QI uses a covariate model and includes 6 covariates in the adjustment model, which the Steering Committee recommended retaining. Future work on this QI should focus on refinement of the adjustment model. The **Little or No Activity (CHSRA)** cross-sectional QI utilizes a restricted resident pool as risk adjustment method, in that residents who are comatose are excluded from the calculation.

Table 6.2.3

Psychosocial QIs Accepted for Further Validation

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Behavior (CHSRA)	VT	0.34		0.20	0.51	0.75*	High/Low Risk Groups (3 Covariates)	This is a prevalence QI of behavioral symptoms affecting others. It is a relatively high prevalence QI measuring an important nursing facility quality issue . Only 5% or less of facilities have no one who has experienced a behavioral decline over the prior 90 day period.
	KS	0.24		0.08	0.42	0.87*		
	NY	0.21		0.09	0.33	0.89*		
Behavior (LTCQ)	VT	0.16	0.15	0.15	0.26	0.22*	Regression Based Adjustment (21 Covariates)	This is an QI of decline in behavioral function over a 3-month period. The average facility rate varies from .07 to .16. Facility rates at the 90th percentile also vary, from 0.12 to 0.26. It is risk-adjusted and is the only longitudinal behavioral measure brought forward for analysis.
	KS	0.08	0.08	0.08	0.15	0.26*		
	NY	0.07	0.07	0.07	0.12	0.37*		
Mood (CHSRA)	VT	0.16		0.06	0.30	0.80	NONE	This is a cross-sectional QI of prevalence of depression in a facility. It is not risk-adjusted, but does provide an indication of mood difficulties among residents in a facility.
	ME	0.08		0.00	0.19	0.77		
	KS	0.07		0.00	0.16	0.80		
	SD	0.04		0.00	0.09	0.73		
	NY	0.03		0.00	0.07	0.71		
Mood (LTCQ)	VT	0.24	0.25	0.12	0.38	0.17	Regression Based Adjustment (17 Covariates)	This is a risk-adjusted QI of decline in mood, including depression, sad mood, and anxiety. It demonstrates moderate interperiod stability. This is the only decline in mood measure which was brought forward for analysis.
	ME	0.17	0.17	0.07	0.29	0.56		
	KS	0.16	0.17	0.06	0.29	0.35		
	SD	0.17	0.17	0.08	0.27	0.51		
	NY	0.10	0.10	0.03	0.17	0.57		

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

1. Cannot perform stability analysis because of a lack of sufficient time periods of data (annual rather than quarterly).

Table 6.2.3

Psychosocial QIs Accepted for Further Validation

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Mood with no treatment (CHSRA)	VT ME KS SD NY	0.09 0.05 0.04 0.02 0.02		0.02 0.00 0.00 0.00 0.00	0.19 0.10 0.12 0.06 0.04	0.80 0.73 0.75 0.65 0.63	NONE	This is a measure of untreated depression or symptoms of distress. It is not risk-adjusted, but it provides a cross-sectional snapshot of untreated depression. It is informative to compare the rates on this QI with the rates of the CHSRA mood QI above. While the preference would be for a risk-adjusted QI, the importance of identifying facilities with inappropriate levels of untreated depression overrides risk adjustment for purposes of bringing forward this QI at this time.
Personal Relationships (LTCQ)	VT KS NY	0.11 0.09 0.04	0.11 0.09 0.03	0.03 0.00 0.00	0.20 0.19 0.09	1. (See below.)	Regression Based Adjustment (6 Covariates)	This measure is based on the MDS annual assessment and is an QI of unsettled relationships. The measure points to conflicts which may reflect difficulties in the social environment of a nursing facility. Despite the measurement problem of being only annual, it is the only QI brought forward which taps into the social environment of a facility.
Little or No Activity	VT KS NY	0.43 0.42 0.32		0.22 0.14 0.07	0.66 0.66 0.61	0.80 0.84 0.96	Restricted Resident Pool (1 covariate)	This is a cross-sectional QI of residents who do not engage is significant social activity on a daily basis. The measure does not take into account activity levels of the residents prior to nursing facility admission. Comatose residents are excluded from the denominator.

For stability, unless noted, correlation is across two adjacent time periods.

*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.

** Percentage of facilities changing 3+ deciles.

* Range of intertime decile correlation.

1. Cannot perform stability analysis because of a lack of sufficient time periods of data (annual rather than quarterly).

6.2.4 Pharmacotherapy QIs Accepted for Further Validation

Two of the four pharmacotherapy QIs developed by CHSRA were accepted for the set of QIs to be provisionally recommended for use by CMS. Both recommended measures are cross-sectional and one of them is casemix-adjusted. The two QIs recommended for further use pertain to antipsychotic drug use and to the use of antianxiety/hypnotic agents. Antipsychotic drugs have been called pharmacological restraints since they are administered to residents with behavioral problems, hallucinations and verbal outbursts associated with brain diseases such as Alzheimer's. The fact that the CHSRA antipsychotic use QI excludes residents with selected psychiatric diagnoses from consideration and then differentiates between high and low risk residents addresses a number of problems that arise because some facilities have historically admitted residents with psychiatric histories.

The antianxiety/hypnotic use QI is not stratified but does exclude residents with selected psychiatric diagnoses. The problematic aspects of antianxiety use as a global QI is that it is not sufficiently precise. While many such drugs can be inappropriate for older persons, the greatest potential for damage lies in receipt of long acting, high dose benzodiazapines. These have been associated with falls and hospitalization for hip fracture, but, depending upon the facility, may only represent a minority of all antianxiety/hypnotic use. Nonetheless, the QI as currently operationalized appears to provide the basis for identifying facilities that may have high use of this more restricted and problematic class of drugs, something that can be checked more completely through on-site inspection.

Table 6.2.4 presents the results of the empirical analyses on the two pharmacotherapy QIs provisionally recommended for use by CMS.

Table 6.2.4
Pharmacotherapy QIs Accepted for Further Validation.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Antipsychotic use (CHSRA)	VT	0.15	*High Risk:* 0.30	0.03	0.14		Restricted Resident Pool	While not a sufficient marker for poor care, anti-psychotic drug use among residents with no psychiatric diagnosis is acknowledged to be inappropriate. This QI drops residents with a psychiatric diagnosis from the resident pool and stratifies the antipsychotic use into high risk and low risk based upon the presence of cognitive or behavioral problems. Facility rates pertaining to residents with high risk are more than twice those of the rates among residents classified as low risk, suggesting construct validity.
			Low Risk: 0.12	0.02	0.12			
	ME	0.14	*High Risk:* 0.27	0.02	0.16			
			Low Risk: 0.08	0.01	0.07			
	KS	0.13	*High Risk:* 0.22	0.02	0.10			
			Low Risk: 0.09	0.01	0.08			
	SD	0.12	*High Risk:* 0.29	0.04	0.13			
			Low Risk: 0.07	0.00	0.07			
	NY	0.14	*High Risk:* 0.34	0.05	0.24			
			Low Risk: 0.10	0.02	0.25			
Antianxiety/ hypnotic use (CHSRA)	VT	0.15		0.02	0.20	0.36	NONE	Because antianxiety/hypnotic agents are linked to falls across all elders, particularly night time use of hypnotics, this QI has content validity. The facility rates at the 90th percentile showed marked variation across states. Some states had rates that were double the rates of others at the 90th percentile. Some of variation may be due to differences in casemix, as this QI is not casemix-adjusted. The rates were stable over successive periods of time.
	ME	0.19		0.03	0.19	0.34		
	KS	0.15		0.02	0.15	0.33		
	SD	0.17		0.04	0.22	0.33		
	NY	0.11		0.03	0.35	0.34		

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

6.3 QIs Rejected Following Initial Analyses

Tables 6.3.1-4. present information on the QIs not recommended for use within each domain. The tables show the distribution of the facility rates by state, the consistency across time periods, the type of risk adjustment, and the main reasons for rejecting the QI.

6.3.1 Functional QIs Rejected Following Initial Analyses

Six functional QIs were not recommended for use: bedfast, cognition, decline in range of motion (2), and functional decline (2). **Bedfast (CHSRA)** was the only functional QI based on cross-sectional information. It looked at the proportion of residents who were in a bed or recliner in their own room for 22 hours or more per day. Bathroom privileges are permitted under this definition; thus, bedfast does not refer to those who are confined to bed or unable to get up on their own. Additionally, bedfast does not represent residents who are totally dependent. Residents could be lifted, placed in a recliner and wheeled into a hallway or common area without being assessed as bedfast. It was rejected primarily because of its low prevalence (0.03-0.06) and because of a lack of risk adjustme nt. In addition, there did not seem to be a strong rationale for a need for an QI in this area.

The **CHSRA cognitive QI** was rejected because it addressed the incidence of new cognitive decline and as such excluded all residents who had any cognitive impairment at baseline (the majority). Not surprisingly therefore, this QI had a low rate within facilities and showed inconsistencies across states and across time.

Two **ROM decline QIs (HCDS and CHSRA)** were rejected. Both looked at decline in range of motio n (ROM) among those who did not have maximal decline at the previous assessment. In addition, the HCDS QI stratified residents into high and low risk groups. High risk was defined as residents who were totally dependent in mobility (bed mobility, transfer, and locomotion). Both were rejected primarily because of the content area. The component MDS items refer to functional joint movement instead of ROM impairment alone. ROM not affecting functional status is not as clinically relevant in most cases. Therefore QIs should target ADL or mobility measures as these are more reliable and directly measure functional independence. Additionally, there is concern that loss of functional ROM may be more often detected in higher functioning residents. Facilities that are more vigilant may be more likely to detect loss of ROM in bedfast or otherwise very dependent residents for whom accurate measurement of ROM is more difficult. Neither ROM QIs could be assessed relative to their stability over time because they required MDS V2.0 items which were only available in Vermont for one quarter.

The two **functional decline QIs (LTCQ and Mukamel)** had similar rates but the covariate models showed inconsistencies in the magnitude of the odds ratios within and across states. A perceived problem with the LTCQ QI was that the regression-based model lacked sufficient clinical content validity. Many of the proposed covariates identified residents with higher functioning who had more opportunity to decline. Given the reservations with the choice of covariates, the Steering Committee was concerned that only 1 point was required to demonstrate change. Although a 1-point change in ADL had great meaning in itself, without good risk adjustment, a 1-point change would be prone to measurement error. An additional problem with the QI developed by Mukamel is that this QI did not exclude residents with maximal decline in function and included variables that were not feasible or not available for current use.

Table 6.3.1
Functional QIs Rejected Following Initial Analyses

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Bedfast (CHSRA)	VT KS NY	0.06 0.05 0.03		0.00 0.00 0.00	0.13 0.12 0.08	0.73 0.80 0.80	NONE	This QI looks at the proportion of residents who were in a bed or recliner in their own room for 22 hours or more per day. The main reason for rejecting this QI is the low prevalence rate and the lack of risk adjustment.
Cognitive (CHSRA)	VT KS NY	0.12 0.09 0.06		0.00 0.00 0.00	0.34 0.17 0.12	0.24 0.36 0.22	Restricted Resident Pool	This QI looks at the incidence of new cognitive impairment as measured by any problem with short-term memory or daily decision-making. All residents with any cognitive problem on the previous assessment were excluded. Therefore the QI applied to a minority of nursing facility residents. Not surprisingly therefore, this QI had a lower incidence rate than the LTCQ cognitive QI within facilities and showed inconsistencies across states and across time.
ROM Decline (HCDS)	VT	0.13	*High Risk* 0.13 *Low Risk* 0.11	0.07 0.04	0.00 0.04	1 (See below)	High/Low Risk	This QI looks at the residents whose functional joint range of (ROM) became more restricted over time. It was rejected primarily because of the content area. Facilities who are more vigilant may be more likely to detect loss of ROM in bedfast or otherwise very dependent residents for whom accurate measurement of ROM is more difficult. They should not be penalized for being more vigilant in reporting.
ROM Decline (CHSRA)	VT	0.12		0.04	0.20	1 (See below)	Restricted Resident Pool	This QI looks at the residents whose functional joint range of (ROM) became more restricted over time. It was rejected primarily because of the content area.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

1. Cannot perform stability analysis because of a lack of sufficient time periods of data (annual rather than quarterly).

Table 6.3.1
Functional QIs Rejected Following Initial Analyses

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
ADL Decline (LTCQ)	VT KS NY	0.27 0.23 0.21	0.27 0.23 0.21	0.13 0.10 0.12	0.42 0.37 0.30	0.01 0.34 0.43	Regression Based Adjustment (8 Covariates)	This QI measures decline in function by a one- level decline in any one of the late loss ADL items (bed mobility, transfer, eating and toileting.). Several covariates in the risk adjustment model appeared to be proxy variables for identifying relatively independent residents who had greater opportunity to decline rather than identifying risk factors for decline. For example it is not clear why being able to establish goals would be a risk factor for functional decline. The questionable clinical content validity was a prime concern. The accepted CHSRA QI was more conservative as it required at least a 2-point change in ADL performance.
ADL Decline (Mukamel)	VT KS NY	0.44 0.28 0.25	0.44 0.29 0.25	0.34 0.14 0.15	0.58 0.43 0.36		Regression Based Adjustment (8 Covariates)	This QI looks at any decline in function as measured by a deterioration in the score of a summary scale that was comprised of the 4 late loss ADL items, each of which was scored 1-5. This QI was not accepted primarily because it did not exclude residents with maximal decline in function. Additionally, all variables originally used in the model were not feasible or not available for current use.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

6.3.2 Clinical Complexity QIs Rejected Following Initial Analyses

Of the twenty-three clinical QIs that were empirically analyzed, the Steering Committee voted to reject 10 QIs (7 cross-sectional; 3 change in status) for use by CMS.

Six of the seven rejected cross-sectional QIs were not risk-adjusted and included: Prevalence of bowel or bladder incontinence without a toileting plan (CHSRA); Fecal Impaction (CHSRA); Dehydration (CHSRA); Feeding Tube (CHSRA); Falls (CHSRA); and Weight Loss (CHSRA). The primary rationale for rejecting these QIs was for lack of risk adjustment in the presence of variation in rates of the measures; the other was due to the questionable inter-rater reliability of the MDS items used to define the QIs (i.e., dehydration; fecal impaction). The other cross-sectional QI, Bladder and Bowel Incontinence without a Toileting Plan (CHSRA), utilizes a restricted resident pool model. This QI was rejected because a large proportion of facilities were lost to analysis from this restriction, thus limiting the QI's applicability. The three other incontinence QIs evaluated were deemed to be sufficient for identifying problems in this area (see Table 6.2.2). Four change measures were also rejected: New Fracture (CHSRA), Physical Restraints (Mukamel; LTCQ) and Pressure Ulcers (Mukamel).

The QI for **Prevalence of bladder and bowel incontinence without a toileting plan (CHSRA)** measures occasional or frequent incontinence of either type without a toileting plan. This measure produces wide variation in the overall rates within and across study states. Lack of risk adjustment makes it difficult to both understand this variation and to make sound comparisons and recommendations for quality improvement. Another key problem is loss of 50 to 68 percent of facilities in the analyses because of the restricted resident pool. This severely limits the applicability of this QI.

The **Fecal Impaction (CHSRA)** QI measures the prevalence of fecal impaction, a symptom which is often related to underlying chronic constipation, inadequate food/fluid intake, or inappropriate use of laxatives. Fecal impaction is often associated with fecal and urinary incontinence and may be a marker of poor quality care. However, the QI as defined does not function adequately for this purpose. The sole MDS item upon which the QI is based has a low inter-rater reliability score of 0.52 and requires a physical or x-ray examination to determine its presence. Therefore, the possibility of ascertainment bias or under-reporting by facilities is great. The overall QI rates vary widely across states and are unstable over adjacent time periods. For these reasons, the SC rejected this QI for use by CMS.

The **Dehydration (CHSRA)** QI measures the prevalence of dehydration, which CHSRA defines by the MDS item "output exceeds input" or by recorded medical diagnosis. Dehydration is a common, serious problem among frail, debilitated nursing facility residents that often leads to hospitalization and poor quality outcomes (e.g., delirium, constipation, falls). It is also a common symptom among persons in their final days of life. Despite its clinical importance, this QI as defined is not an adequate QI of overall nursing facility quality. First, although it functions as a sentinel event, one of the primary MDS items upon which it is based has very poor inter-rater reliability, making it unstable. The other item is based on a physician's diagnosis. When recorded as a diagnosis it is most likely accurate; when not, the problem may be underestimated. Second, the item may be "gameable" by facilities (i.e., because it is not readily observable, it may not be recorded by staff) and would be difficult to monitor by long-term care surveyors. Third, because adequate treatment often precipitates hospitalization, the risk of censoring bias is high. Finally, this QI lacks risk adjustment, placing facilities with a casemix of persons at higher risk for dehydration (e.g., medically ill, terminally ill) at greater disadvantage in a regulatory system. Lack of casemix adjustment may also inadvertently prompt such facilities to inappropriately hydrate persons approaching death.

The **Restraints QI (Mukamel)** is risk-adjusted but was difficult to estimate using MDS data because it was originally developed using the PRI assessment tool used by New York facilities for Medicaid reimbursement. All PRI data collection was different in scope, timing and measurement definition from that of MDS data. An attempt was made to reproduce PRI mobility and transfer function covariates but a comparable match between the PRI and MDS behavioral covariates was not possible. Despite these difficulties the QI functioned well statistically. The problem the SC had with the QI was primarily conceptual and related to the nature of the covariate adjustment. Mukamel's model is based solely on physical performance measures of mobility and transfer as well as behavioral symptoms; and could be interpreted as iatrogenesis. The SC deemed the covariates inappropriate for use in the QI.

The **LTCQ Restraints** QI, which measures the incidence of new physical restraint, is adjusted with 24 covariates. Although the SC agreed that new placement of a restraint on a nursing facility resident is an important quality event that deserves monitoring, the QI was deemed to be overadjusted.

The **Tube Feeding (CHSRA)** QI measures the prevalence of feeding tube use. It excludes admission and readmission assessments and thus reflects use of feeding tubes as a function of care in the nursing facility. There is wide variation in overall rates across states with one state having approximately three times the use of other study states. There is extremely wide variation across states in the number of facilities having zero use of feeding tubes. Likewise, there is variation in percentile rankings of facilities within and across states. Because this QI is not risk-adjusted it is not possible to begin to understand the reason for such variation. This QI as defined is not adequate for estimating and comparing the quality of facilities with regard to feeding tube practices.

The **Falls (CHSRA)** QI measures the prevalence of falls in the past 30 days. Overall rates are stable across states but there is wide variation in the percentile rankings of facilities within states. Because this QI is not risk-adjusted it is not possible to begin to understand the reason for such variation. This QI as defined is not adequate for estimating and comparing the quality of facilities with regard to prevalence of falls.

The **New Fracture (CHSRA)** QI measures the incidence of new fractures over the last quarter. Overall rates are very low and stable across states, but there is some variation in the percentile rankings across facilities within states. Because this QI is not risk-adjusted it is not possible to understand the reason for the variation across percentile rankings within states and therefore, as defined, is not adequate for estimating and comparing the quality of facilities with regard to new fracture incidence. The **Mukamel Pressure Ulcer QI** utilizes a 13-covariate adjusted model to measure incidence or worsening of ulcers. This QI is stable over time; findings show that facilities with low rates are likely to continue to have low rates over the next quarter. Although the mean QI rates and adjustment findings are comparable to the accepted LTCQ Pressure Ulcer QI, this covariate model is more difficult to follow (e.g., includes squared RUGs scores and dummy variables for various functional levels) and is less inclusive of relevant findings in recent literature.

Weight loss is a common, morbid problem in nursing facility residents. The overall mean rates for the **CHSRA Weight Loss** QI reflect that on average ten percent of residents across study states have lost five percent or more weight in the past 30 days or ten percent in the last six months. In many cases this level of weight loss is preventable, and under facility control. In other cases, it is an expected part of the disease process (e.g., terminal cancer) over which residents and their families have the right to exercise their preferences over care interventions (e.g., accepting/refusing enteral support). Although overall mean rates of weight loss are stable across states, there is wide variation in the facility percentile rankings within states. Lack of casemix adjustment makes it difficult to interpret these variations or use for making comparisons; and it may punish facilities who care for a large casemix of debilitated, terminally ill persons. It also may inadvertently prompt such facilities to inappropriately artificially feed persons approaching death.

Table 6.3.2
Clinical Complexity QIs Rejected Following Initial Analyses.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Prevalence of bladder and bowel incontinence without a toileting plan (CHSRA)	VT ME KS NY SD	0.33 0.53 0.36 0.36 0.38		0.09 0.19 0.04 0.11 0.10	0.72 0.94 0.89 0.74 0.69		Restricted Resident Pool	This prevalence QI for the presence of occasional or frequent incontinence without a toileting plan produces wide variation in the overall rates across study states. Likewise, there is wide variation across percentile rankings of facilities within and across states. Lack of risk adjustment makes it difficult to understand this variation for making sound comparisons and recommendations for quality improvement.
Fecal Impaction (CHSRA)	VT ME KS NY SD	0.01 0.00 0.01 0.01 0.00		0.00 0.00 0.00 0.00 0.00	0.03 0.03 0.02 0.02 0.01	0.33 0.27 0.22 0.35 0.15	NONE	The symptom of fecal impaction may be a marker of poor quality care. However, the QI as defined functions poorly for this purpose. The sole MDS item upon which the QI is based has an inter-rater reliability of 0.52 (acceptable but low). The overall QI rates vary across states, and between 48.8% (VT) and 84.8% (SD) of facilities have zero prevalence. The stability of the QI over two adjacent quarters is also weak. For these reasons, the SC rejected this QI for future use.
Dehydration (CHSRA)	VT KS NY	0.01 0.01 0.02		0.00 0.00 0.00	0.02 0.02 0.05	0.52 0.41 0.52	NONE	Dehydration is a problem among frail, debilitated nursing facility residents that often leads to hospitalization and poor quality outcomes (e.g., delirium, constipation, falls). It is also a common symptom among persons in their final days of life. Despite its clinical importance, this QI as defined is a poor QI of quality. This QI lacks risk adjustment, placing facilities with a casemix of persons at higher risk for dehydration (e.g., medically ill; terminally ill) at greater disadvantage in a regulatory system. Lack of casemix adjustment may also inadvertently prompt such facilities to inappropriately hydrate persons approaching death.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

Table 6.3.2
Clinical Complexity QIs Rejected Following Initial Analyses.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Restraints (Mukamel)	VT KS NY	0.46 0.54 0.57	0.45 0.59 0.53	0.20 0.37 0.15	0.70 0.82 0.83	0.88 0.94 0.96	Regression Based Adjustment (6 Covariates)	This covariate-adjusted QI measures the incidence of new physical restraint, which if defined adequately could be used to measure sentinel events as the overall raw and adjusted rates across study states is very low. Although the SC agreed that placing a nursing facility resident in a physical restraint is an important quality event that deserves monitoring, the covariates were deemed to be inappropriate for use in the QI.
Restraints (LTCQ)	VT KS NY	0.01 0.01 0.01	0.01 0.01 0.01	0.00 0.00 0.00	0.03 0.03 0.03	0.26** 0.30** 0.26**	Regression Based Adjustment (24 Covariates)	This covariate-adjusted QI measures the incidence of new physical restraint, which if defined adequately could be used to measure sentinel events as the overall raw and adjusted rates across study states is very low. Although the SC agreed that placing a nursing facility resident in a physical restraint is an important quality event that deserves monitoring, the QI was deemed to be overadjusted.
Tube Feeding (CHSRA)	VT ME KS NY	0.02 0.02 0.02 0.08		0.00 0.00 0.00 0.01	0.06 0.05 0.05 0.15	0.91 0.94 0.78	NONE	This QI measures the prevalence of feeding tubes. It excludes admission and readmission assessments and thus reflects use of feeding tubes as a function of care in the nursing facility. There is wide variation in overall rates, and extremely wide variation across states in the number of facilities having zero use of feeding tubes (6.7% in NY - 44% in KS). Likewise, there is variation in percentile rankings of facilities within and across states. Lack of risk adjustment makes it difficult to understand this variation for making sound comparisons and recommendations for quality improvement. This QI as defined is not adequate for estimating and comparing the quality of facilities with regard to feeding tube practices.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

Table 6.3.2
Clinical Complexity QIs Rejected Following Initial Analyses.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Falls (CHSRA)	VT KS NY	0.17 0.16 0.13		0.09 0.07 0.07	0.28 0.22 0.21	0.52 0.49 0.79	NONE	This QI measures the prevalence of falls in the past 30 days. Overall rates are stable across states but there is wide variation in the percentile rankings of facilities within states. Lack of risk adjustment makes it difficult to understand this variation for making sound comparisons and recommendations for quality improvement. This QI as defined is not adequate for estimating and comparing the quality of facilities with regard to prevalence of falls .
New Fracture (CHSRA)	VT KS NY	 0.02 0.01		0.00 0.00 0.00	0.05 0.05 0.03	-0.01 0.00 0.16	NONE	This QI measures the incidence of new fractures over the last quarter. Overall rates are very low in all states, but there is some variation in the range of rates across states. Lack of risk adjustment makes it difficult to understand this variation for making sound comparisons and recommendations for quality improvement. Although a QI of this type could be used to measure sentinel events, because rates approach 5% at the 90th percentile in some states, it would be a more useful QI if it were risk-adjusted.
Pressure Ulcers (Mukamel)	VT KS NY	0.05 0.04 0.05	0.06 0.04 0.06	0.00 0.00 0.02	0.10 0.09 0.10	0.26 0.26 0.44	Regression Based Adjustment (13 Covariates)	This covariate-adjusted QI measures pressure ulcer incidence or worsening. This QI is stable over time; findings show that facilities with low rates are likely to continue to have low rates over the next quarter. Because the covariate model is difficult to follow, and a risk-adjusted pressure QI was already accepted (LTCQ), this QI was rejected.

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

Table 6.3.2
Clinical Complexity QIs Rejected Following Initial Analyses.

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Weight Loss (CHSRA)	VT	0.11		0.03	0.18	0.28	NONE	Weight loss is a common, morbid problem in nursing facility residents; the overall mean rates for this prevalence QI reflect that on average 10% of residents across study states have lost 5% or more in the past 30 days or 10% in the last 6 months. In many cases this level of weight loss is preventable, and under facility control. In other cases, it is an expected part of the disease process (e.g., terminal cancer) over which residents and their families have the right to exercise their preferences over care interventions (e.g., accepting/refusing enteral support). However, lack of covariate adjustment makes this QI unusable as an indicator of nursing facility quality.
	KS	0.10		0.03	0.18	0.39		
	NY	0.09		0.03	0.17	0.36		

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

6.3.3 Psychosocial QIs Rejected Following Initial Analyses

There were no psychosocial QIs that were rejected following initial analyses.

6.3.4 Pharmacotherapy QIs Rejected Following Initial Analyses

Two of the four QIs focusing on prescribing patterns were rejected by the Steering Committee, albeit for very different reasons. One of these, **Receipt of antianxiety/hypnotic agents twice a week or more (CHSRA)** mirrors the accepted QI pertaining to antianxiety drug use very closely, meaning that it adds little to the information already contained in the accepted QI. The other rejected QI, **Use of 9 or more medications (CHSRA)**, is based upon the assumption that the more drugs one is taking the greater the risk of an adverse event. While the literature certainly confirms this rather rudimentary observation, by and large this research has been based upon a count of prescription drugs rather than all agents, including those that would be "over the counter" drugs for community dwelling elders. Thus, in the nursing facility context, not only is this QI imprecise, it is also inflated since it includes many agents that do not increase the risk of adverse events, either directly or indirectly.

Table 6.3.4
Pharmacotherapy QIs Rejected Following Initial Analyses

QI	State	Raw Rate	Adjusted Rate	***Rate At:		Stability	Risk Adj Method	Primary Rationale for Decision
		Mean	Mean	10th %	90th %			
Use of 9 or more medications (CHSRA)	VT	0.26		0.04	0.35	0.32	NONE	This QI measures whether residents take over 9 pharmaceutical agents over a 7-day period. Since this count includes non-prescription agents and there is limited evidence of harm associated with multiple drug use, content validity of the measure is limited.
	ME	0.34		0.08	0.42	0.33		
	KS	0.31		0.08	0.32	0.34		
	NY	0.24		0.05	0.66	0.33		
	SD	0.33		0.12	0.42	0.34		
Hypnotic use >2 (CHSRA)	VT	0.15		0.02	0.20		NONE	This QI measures the proportion of residents taking an antianxiety/hypnotic agent at least twice in a week. Since virtually all residents taking these agents once a day also take them twice a day, this QI is redundant and adds no further information on the quality of medication management.
	ME	0.19		0.03	0.29			
	KS	0.15		0.02	0.15			
	SD	0.17		0.04	0.22			
	NY	0.11		0.03	0.35			

For stability, unless noted, correlation is across two adjacent time periods.
*** Values represent adjusted rates. If rates were not adjusted, raw rates were used.
** Percentage of facilities changing 3+ deciles.
* Range of intertime decile correlation.

7.0 Development of an Adjustment Method for Existing QIs

Based on the conceptual and descriptive analyses presented in Chapter 6 of this report, the project team identified 26 existing QIs as promising candidates for use by CMS. However, additional empirical analyses, described in detail in Technical Appendix 1, suggested that the QIs in their current form might lead to incorrect assignment of some facilities as providing poor care. The main concern was that the existing QIs were not adequately adjusted for differences in casemix and in assessment accuracy across facilities, rendering them vulnerable to selection and measurement (ascertainment) bias. Consequently, the project team decided that all 26 QIs were suitable for nursing facilities in their internal quality improvement efforts. However, without additional refinement, none of them could be used for public reporting to consumers or purchasers, and only a few of them for the survey process.

This chapter describes the development of a new adjustment methodology for the existing QIs. As a first step, the project team modified the covariate structure of the LTCQ QIs by removing some risk adjusters with questionable clinical relevance and by dropping so-called service variables. Those variables reflect services that are to some degree at the discretion of the facility (e.g., bedrails or urinary catheters). The concern is that including such variables in risk adjustment models provides an incentive for facilities to use those services more so that their patterns of care would be distorted and the actual severity of their casemix would be overstated.

This chapter first reviews the goals of casemix adjustment, the method of adjustment used previously and some of the computational and conceptual limitations of this method. Then, a new method of adjustment is presented and illustrated with an example QI. We introduce the concept of a 'facility admission profile' (FAP) variable and the importance of including such a measure in the casemix adjustment. Finally, descriptive statistics for the QIs previously accepted for use under the new casemix adjustment procedure are presented, along with a summary of the Steering Committee's recommendations on applicability of each QI for internal quality improvement efforts, surveyors, and for public dissemination.

7.1 Rationale for Casemix Adjustment

Individual residents face differential risk for specific adverse events given their varying health and functional status. Some of these predisposing characteristics increase the risk of adverse outcomes independent of quality of care. Thus, facility populations vary in the level of overall health and functional impairment due to varying admission practices or differences in discharge practices. Consequently, adjusting for casemix is a key aspect of any outcome and QI measurement. In addition, one has to keep in mind that QIs are being constructed from secondary data so that differences in recording practices across facilities could lead to different reported QI scores for facilities which have in reality the same number of events. As long as differences in casemix and reporting are random, i.e., unrelated to quality of care, they would only decrease the precision with which a QI can measure quality. However, since it is likely that excellent facilities both attract sicker residents and report more accurately, unadjusted QIs might be biased against better facilities and reflect misleading judgments about quality of care.

7.1.1 Alternative Casemix Adjustment Models

We investigated different types of casemix adjustment models. These included:

(a) a dichotomous risk model;
(b) direct adjustment using risk strata; and
(c) regression-based methods.

Dichotomous casemix adjustment models are easy to implement and the meaning of obtained results are easy to communicate. However, such methods may not be adequate to describe a meaningful array of important resident characteristics on risk of outcome. The dichotomous risk-model is typical of the CHSRA QI's. The strength of this model is the fact that it does identify a sub-group of persons at higher risk of an adverse outcome; its weakness is that it fails to capture the observed increase in risk associated with increased severity. For example, one-year pressure ulcer incidence ranges from less than 5 percent to nearly 40 percent as a function of the number of different risk factors present in a resident. However, only 20 percent of nursing home residents have no (0) risk factors. Simply dividing residents into those who are and are not at risk can miss significant variation in risk and its relationship to the outcome.

Direct standardization involves stratifying residents into risk strata defined by a number of characteristics thought to be important risk adjusters, and computing a facility-level prevalence with each strata contributing equally to the overall total across facilities. The strength of this approach is that each adjusted QI is literally that value of the QI that would be observed if all facilities had the same casemix, and that adjusted QI values fall within a logical range (never greater than 100 percent). This method is limited, however, in that the number of included resident characteristics is limited, for as the number of characteristics increases, so too do the number of strata. As the number of strata increases, the number of residents within a facility populating those strata becomes sparse and the resulting adjusted QI can be highly unstable.

For those reasons, the project team adopted a regression-based adjustment method. This method consists of two steps. The first involves estimating a logistic regression model in which each resident's risk of experiencing the QI-defining event is predicted by resident-level and facility-level variables. The resident-level variables reflect clinical characteristics related to the outcome of interest, and were derived from the original covariate list of the respective QI developer. As not all of the 26 recommended QIs had risk adjustment covariates, some indicators could not be adjusted for resident-level factors. However, all QIs were adjusted by including a facility-level variable in the regression model in order to capture unaccounted differences in casemix and assessment acumen. This variable, called "Facility Admission Profile", reflects the level of severity or prevalence of a specific problem among the residents admitted to a given facility. This adjustment is important, since it accounts not only for differential risk of health and functional decline of typical residents in the facility, but also adjusts for differences between facilities in terms of the ability of staff to observe and record clinical features difficult to monitor and assess or the sometimes subtle characteristics of nursing home residents. The operational definitions of the admission profile variable are summarized in Table 7.1.1.

Table 7.1.1

Accepted Quality Indicators and Specified Facility Admission Profile Variables.

Quality Indicator	**Facility Admission Profile Variable**	**Quality Indicator**	**Facility Admission Profile Variable**
ADL		*Falls*	
ADL decline	ADL Long Scale (Morris et al., 1999)	Falls change	Prevalence of falls
Behavior		*Mobility*	
Behavior	Prevalence of verbally, physically abusive, or inappropriate behavior	Mobility Change	Locomotion (G1ae)
Behavior high risk		*Mood*	
Behavior low risk		Depression without treatment	DRS (Burrows et al., 2000)
Behavior change	Sum of behavioral problems (E4a-e)	Depression with treatment	
Catheter		Depressed mood change	
Indwelling urinary catheter	Prevalence of indwelling catheter (H3d)	*Nutrition*	
Catheter		Tube feeding prevalence	Prevalence of tube feeding (K5b>0)
Continence		*Pain*	
Bowel & bladder incontinence	Prevalence of either bowel or bladder incontinence (H1a>2 or H1b>2)	Pain change	Level of pain symptoms (J2b)
Bowel & bladder incontinence high risk		*Pressure Ulcer*	
Bowel & bladder incontinence low risk		Pressure ulcer prevalence	Prevalence of pressure ulcer (M2a>0)
Bowel incontinence change	Bowel continence (H1a)	Pressure ulcer prevalence high risk	
Bladder incontinence change	Bladder continence (H1b)	Pressure ulcer prevalence low risk	
Urinary tract infection prevalence	Prevalence of urinary tract infection (I2j)	Pressure ulcer change	
Cognition		*Restraints*	
Cognition change	CPS (Morris et al., 1994)	Restraint prevalence	Prev. restraint use (P4c-e equal to 2)
Communication		*Social*	
Communication change	Sum of items C4 and C6	Personal relationships change	Sum of personal relation items (F2a-d)
Drugs		Little or no activities	Prevalence of no/little activities (N2=2|3)
Antipsychotic prevalence	Prevalence of antipsychotic use O4a>0	*Weight*	
Antipsychotic prevalence high risk		Weight loss	Weight loss (K3a>0)
Antipsychotic prevalence low risk			
Anxiolytic/hypnotic prevalence	Prevalence of anxiolytic/hypnotic use		

CPS - Cognitive Performance Scale
DRS - Depression Rating Scale

After running the logistic regression models, one retrieves a facility's expected event rate by summing the predicted probabilities of all residents for that facility. The ratio of the actually observed over this predicted event rate reflects whether a facility has more (ratio>1.0) or fewer (ratio <1.0) events than predicted by the model, using all the other facilities in the sample as reference.

As ratios are more difficult to interpret intuitively, the project team decided to convert this ratio into an adjusted rate. This can be achieved by multiplying the ratio by the mean event rate in the sample, usually all the facilities in a state. This procedure, which will be referred to as *indirect linear adjustment*, can be summarized with the equation

$$\hat{p}_j = \left(\frac{p_j}{P_j}\right)\overline{P}_j \qquad (1)$$

where $\hat{P}j$ is the facility-level adjusted rate for the jth event, p_j is the observed facility-level rate, P_j is the expected facility-level rate, and $\overline{P}j$ is the state-level mean event rate.

The indirect linear adjustment method is related to the epidemiologic adjustment procedure called indirect adjustment (Kahn & Sempos, 1989), and has a number of appealing features.

Firstly, the resulting adjusted QI $\hat{P}j$ had a state-level mean that was essentially identical to the unadjusted state-level mean. Secondly, facilities that had a QI prevalence greater than what would be expected given their casemix would have adjusted QI values above the state mean. Likewise, facilities that had an observed QI prevalence lower than expected given their casemix would have an adjusted QI value lower than the state-level mean. Facilities with an unadjusted zero prevalence will have an adjusted zero prevalence.

7.1.2 Indirect Standardization with Non-Linear Transformation

Problems with the method of adjustment outlined in equation (1) above is that some facilities, in rare circumstances, could have an adjusted QI rate of greater than 1. In other words, the adjusted prevalence of the QI in the facility was more than 100 percent. This would occur when there were wide discrepancies between the observed and expected facility rate, and may occur when the logistic model used to generate the predicted models fit the observed data poorly. An easy solution to the problem of inadmissible adjusted QI values is to transform the probabilities to their normal deviates prior to arithmetic manipulation. This procedure, which will be referred to *as indirect standardization with probit adjustment*, or *probit adjustment* can be summarized with

$$\hat{p}_j = \begin{cases} \phi\{[\phi^{-1}(p_j) - \phi^{-1}(P_j)] + \phi^{-1}(\bar{P}_j)\} \text{ if } 0<p_j<1 \\ 0 \text{ if } p_j=0 \\ 1 \text{ if } p_j=1 \end{cases} \tag{2}$$

The probit transformation (N^{-1}(.)) involves replacing the observed proportion with the corresponding normal equivalent deviate, or z-score, and performing the computations and scaling with these z-scores or probits. Then, the resulting figure is back-transformed from the standard normal to a probability (N(.)). The transformation removes the possibility of observing facilities with a prevalence of greater than 1, and reduces the uneven variance of adjusted scores as they deviate from zero prevalence.

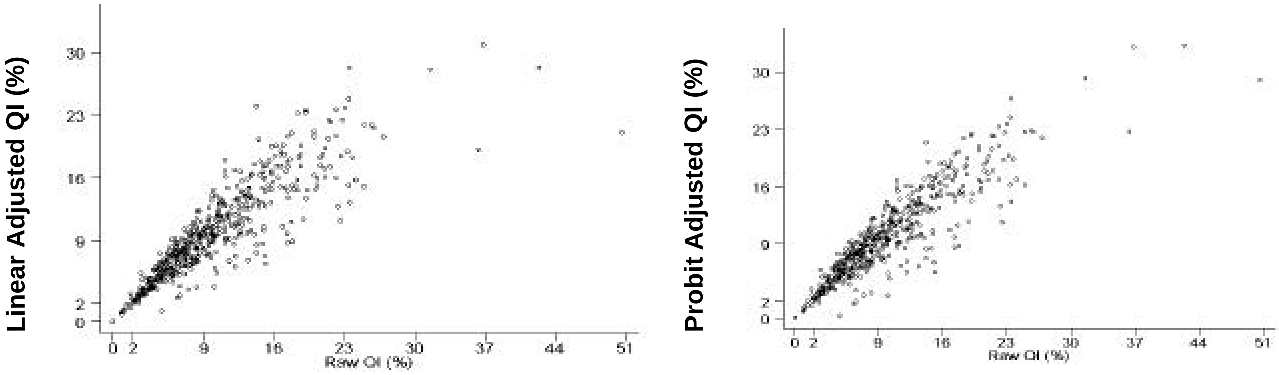
The goal of the adjustment could be accomplished by converting the proportions (p) to logits or log

$$\hat{p}_j = \begin{cases} \left[1+e^{\left(-1\times\left\{\left[ln\left(\frac{p_j}{1-p_j}\right)-ln\left(\frac{P_j}{1-P_j}\right)\right]+ln\left(\frac{\bar{P}_j}{1-\bar{P}_j}\right)\right\}\right)}\right]^{-1} \text{ if } 0<p_j<1 \\ 0 \text{ if } p_j=0 \\ 1 \text{ if } p_j=1 \end{cases} \tag{3}$$

odds (*ln*(p/(1-p))) instead of probits, as shown in equation 3, where *ln* implies the natural logarithm, and *e* exponentiation. From a mathematical perspective, it is immaterial if a probit or logit transformation is used in computing the adjusted QI. The purpose of the transformation is to move the obtained proportions (raw, expected, overall mean prevalence) from a scale bounded by 0 and 1 to a scale without bounds. Arithmetic manipulations are performed in the transformed scale, and the result back-transformed to a 0 to 1 scale. This process of standardization is very different from the process of risk modeling used to estimate the facility expected rate given casemix and admission profile.

Example: Bowel Continence Change. The suitability of the indirect standardization with probit adjustment procedure is exemplified by considering the LTCQ QI capturing quarterly change in bowel continence. The two adjustment methods are illustrated in Figure 7.1. This QI demonstrates the particular problem of the linear adjustment method. One facility has a linear-adjusted QI prevalence of 108 percent. This facility has an observed (unadjusted) prevalence of 60 percent; the probit-adjusted prevalence is 57 percent. Note that the statewide prevalence of this QI is about 19 percent (unadjusted). Notice the fan-spread shape does not characterize the probit-adjusted to the sample degree as the linear-adjusted QI. The absolute magnitude of the difference between the observed and adjusted QI is strongly related to the observed value for the linear-adjusted QI (r=0.41) and only weakly related to the magnitude using the probit-adjusted QI (r=0.12).

Figure 7.1
Scatter-plot of observed (x-axis) and linear-adjusted (y-axis, panel A) and probit-adjusted (y-axis, panel B) LTCQ QI for Quarterly change in bowel continence. Massachusetts repository data (N=524 facilities).



7.2 Issues

There are a number of issues relative to casemix adjustment of quality indicators that deserve comment. These include: 1) the size of the denominator, or number of applicable residents in a facility for a particular QI; 2) assumptions regarding the constancy of casemix adjustment models and covariate lists across state and over time; and 3) the implications of including facility-level effects in resident-level models used to compute expected QI values.

7.2.1 Size of the Denominator

The numerical estimate provided by an unadjusted QI is a proportion, and as such is defined by a numerator (number of residents with the characteristic) and a denominator (number of residents to which the QI is relevant). One of the features important to the stability of a QI is the number of residents in the denominator. If the number of residents in the denominator of the QI is small, there may be wide temporal variability in the obtained QI not directly due to fluctuations in the quality of care provided by the facility, but rather due to the variability in estimation.

The project team originally coded adjusted QIs for facilities only when the number of residents in a given facility to whom the QI is relevant is equal to or greater than 20. In other words if the denominator for a QI is less than 20 in a facility, that facility does not obtain a value for the adjusted QI. This procedure should not be viewed as firm recommendation by the project team. Rather, it is viewed as the best approach at this point. However, alternatives to this procedure have been investigated, including expanding denominators by pooling data across periods of observation within a facility.

7.2.2 Model and Covariate Assumptions

An important feature of our method of casemix adjustment is that it involves a modeling process. In other words, the resident-level risk of satisfying the QI definition is modeled as a function of a specified set of QI-specific covariates. Our method makes no assumption of the constancy of the parameter estimates for this model across state and over time. This means that two facilities with identical residents and identical raw (unadjusted) QI scores could have slightly different adjusted QI scores if they were located in different states, or a facility with identical and unchanging residents could have a slightly different adjusted QI score over time. These differences across state and time are driven by state and temporal variation in the predictive ability of the QI-specific covariates and changes in the overall prevalence of the QI over time and across state.

Considerable simplicity in coding QIs could be achieved by assuming constancy of parameter values in the regression models across time and across state. Under such an assumption, QIs could be coded and computed outside of a formal statistical modeling system (e.g., spreadsheet or database system). However, the assumption of constancy of parameter estimates, and even selected covariates, is not fully supported by empirical observations of repository data. If such a strategy were adopted, it would be essential to periodically re-evaluate the suitability of covariates and the adequacy of the regression and casemix adjustment models, and update accordingly.

On the other hand, our method does involve the assumption of temporal and geographic constancy of the appropriate resident characteristics used in the modeling step. This assumption is warranted given the subtleties of choosing appropriate resident characteristics for use in casemix adjustment. Appropriate risk adjusters are resident characteristics thought to be related to individual risk of satisfying the definition of the QI but not a direct consequence of quality of care provided by the facility.

7.2.3 Facility-level Effects in Casemix Adjustment Models

An important limitation of including the facility-level variable in the resident-level model is that this is not entirely correct from a statistical modeling perspective. Technical Appendix 2 briefly describes more appropriate multi-level, or hierarchical, estimation routines using specialized computer software. However, the current state of the art precludes the use of such algorithms in the automated adjusted QI generation procedure. Furthermore, based upon a limited comparison of the results of multi-level and resident level regression models, we believe that the inclusion of facility-level effects in the resident-level regression models will produce relatively un-biased estimates of covariate parameters and therefore un-biased estimates of expected QI rates in the facility in spite of the fact that they will produce biased estimates of the covariate standard errors. Because biased estimates of covariate standard errors will not influence the adjusted QI estimate, we believe the method of including the facility-level effect in the resident-level model is appropriate. This is an area of continued investigation by the project team.

7.3 Results of Casemix Modeling

In this section, results of applying the Facility Admission Profile adjustor to the 26 accepted QIs are presented. A file with MDS V2.0 repository records prepared by CMS as a single "flat-file" dataset of all MDS record types was used to create two state-specific analytic files, which in turn were used to code and compute the QIs. [The data represent 554 facilities from the state of Massachusetts covering the final of four quarters spanning October 1998 through September 1999.] The first of two working files contains resident-level data. These files were organized so that each facility resident represents one record in the file, and this resident could have one MDS assessment in each of four quarters of calendar time. The second of two working files contains facility-level data reflecting specific characteristics of residents admitted to that facility over a nine-month period (FAP variables). These two working files are used to generate a single facility-level file containing, for each facility in the state, an array of QI scores for each quarter.

The tables below present mean QI rates (Table 7.3.1), the relationship of covariates and QIs at the resident level (the results of the resident-level casemix adjustment models, Table 7.3.2) and the correlation of the unadjusted QI with the adjusted QI and the mean facility covariate score (Table 7.3.3) as well as the facility admission profile score (Table 7.3.4).

It is important to point out an analytic detail pertaining to the results shown in Table 7.3.2. These tables present, for each QI, the risk adjusters (covariates), their prevalence among residents in the quarter of observation, their individual (crude) association with the QI, and their multivariable adjusted association with the QI, holding constant the effect of the other resident-level QIs and the facility admission profile covariate. The measure of association provided is the odds ratio (OR), which can be interpreted as the increase in odds of the outcome (having the QI-defining event) *per unit increase in the specific covariate.*

For the resident-level covariates (which can be symbolized with x), a *per unit* increase indicates the increase in odds when comparing residents without the covariate (x=0) to those with the covariate (x=1). However, facility admission profile variables are on an idiosyncratic scale. For example, the Tube Feeding facility admission profile variable may take any value between 0 and 1, whereas the Behavior facility admission profile variable may take on any value between 0 and 15. In order that the magnitude of the influence of the facility admission profile variable be interpretable and consistent across QIs, the odds ratios presented in Table 7.3.2 have been modified to display the increase in odds of the QI *per standard deviation (F) increase* in the generally continuously distributed facility admission profile variable. In summary, the odds ratios for the crude and multivariable adjusted resident-level covariates provide the exponentiated logistic regression coefficient ($e^{\$}$), whereas the odds ratios for the facility admission profile provide a modified exponentiated logistic regression coefficient ($e^{\$F}$).

For example, Table 7.3.2 displays results from the resident-level casemix adjustment logistic regression model for the LTCQ Behavior Change QI. Four resident-level covariates are included: ability to communicate with speech; moderate or severe impairment in cognitive skills for daily decision making; presence of verbally abusive behavior; and presence of motor agitation. The "Prevalence" column displays how many per 100 residents in the state had that particular characteristic. The "Crude OR" column displays the increase (or decrease) in odds of individuals with the covariate relative to those without the covariate. For example, the odds of a resident with cognitive impairment having a decline in behavior problems is nearly twice (OR=1.9) that of residents without cognitive impairment. The "Adjusted OR" column provides the effect of the covariate holding constant the influence of the other covariates listed, including the facility admission profile effect.

The OR associated with the admission profile effect (OR=1.2) does *not* represent the increase in odds of the QI *per unit* increase in facility admission profile covariate, but instead provides the increase in odds *per standard deviation* increase in the facility admission profile covariate. The mean sum of behavior items among facility admissions over a nine-month period is 0.62 and has a standard deviation (*F*) of 0.61. That is, the facility-level mean (±F) for this admission profile variable is 0.62±0.61. Thus, our model implies that the odds of displaying the QI for a resident from a facility with a value on the facility admission profile variable of 1.84, two standard deviations above the mean, is about 1.4 times that of a resident from a facility with the state average facility admission profile score.

A number of interesting results displayed in the tables are worth mentioning. First, the distribution of QIs is broad, ranging from rare (one percent, personal relationships change) to relatively common (56 percent, bladder & bowel incontinence among low risk residents). Table 7.3.2 highlights the parsimonious casemix adjustment models, and provides evidence that the resident-level risk factors do indeed explain variability in the likelihood of satisfying the QI definition.

Table 7.3.1
Resident-level QI Base Rates (Prevalence). Massachusetts MDS Repository Data. [n = 554 facilities]

		Prevalence (%)
QI		
	ADL	
	ADL decline (CHSRA)	15
	Behavior	
	Behavior (CHSRA)	18
	Behavior high risk (CHSRA)	22
	Behavior low risk (CHSRA)	7
	Behavior change (LTCQ)	9
	Catheter	
	Indwelling urinary catheter (LTCQ)	2
	Catheter (CHSRA)	6
	Continence	
	Bowel & bladder incontinence (CHSRA)	41
	Bowel & bladder incontinence high risk (CHSRA)	6
	Bowel & bladder incontinence low risk (CHSRA)	56
	Bowel incontinence change (LTCQ)	17
	Bladder incontinence change (LTCQ)	19
	Urinary tract infection prevalence (CHSRA)	10
	Cognition	
	Cognition change (LTCQ)	13
	Communication	
	Communication change (LTCQ)	11
	Drugs	
	Antipsychotic prevalence (CHSRA)	20
	Antipsychotic prevalence high risk (CHSRA)	39
	Antipsychotic prevalence low risk (CHSRA)	18
	Anxiolytic/hypnotic prevalence (CHSRA)	18
	Falls	
	Falls change (LTCQ)	14
	Mobility	
	Mobility Change (LTCQ)	13
	Mood	
	Depression without treatment (CHSRA)	2
	Depression with treatment (CHSRA)	2
	Depressed mood change (LTCQ)	14
	Nutrition	
	Tube feeding prevalence (Ramsey)	4
	Pain	
	Pain change (LTCQ)	9
	Pressure Ulcer	
	Pressure ulcer prevalence (CHSRA)	7
	Pressure ulcer prevalence high risk (CHSRA)	16
	Pressure ulcer prevalence low risk (CHSRA)	3
	Pressure ulcer change (LTCQ)	6
	Restraints	
	Restraint prevalence (CHSRA)	6
	Social	
	Personal relationships change (LTCQ)	1
	Little or no activities (CHSRA)	23
	Weight	
	Weight loss (LTCQ)	7

Table 7.3.2
Association of Covariates and QI Prevalence at Resident Level, Massachusetts MDS Repository Data [n=554 facilities]

	Prevalence (per 100 res)	Crude OR	Adjusted OR[1] (95% CI)
QI			
ADL			
ADL Decline (CHSRA)			
facility admission profile			1.0 (1.0, 1.1)
Behavior			
Behavior (CHSRA)			
facility admission profile			1.5 (1.5, 1.5)
Behavior High Risk (CHSRA)			
facility admission profile			1.3 (1.3, 1.4)
Behavior Low Risk (CHSRA)			
facility admission profile			2.0 (1.9, 2.2)
Behavior Change (LTCQ)			
mode of expression	94	1.8	2.2 (1.7, 2.8)
cognitive skills for daily living	56	1.9	1.9 (1.7, 2.1)
verbally abusive behavior	09	1.5	1.3 (1.1, 1.5)
motor agitation	12	1.7	1.5 (1.3, 1.7)
facility admission profile			1.2 (1.2, 1.3)
Catheter			
Indwelling urinary catheter (LTCQ)			
bowel incontinence	28	1.4	1.2 (1.1, 1.4)
pressure ulcer	10	2.2	1.9 (1.6, 2.3)
feeding tube	04	2.0	1.5 (1.2, 2.0)
facility admission profile			1.4 (1.3, 1.5)
Catheter (CHSRA)			
facility admission profile			1.5 (1.4, 1.6)
Continence			
Bowel & bladder incontinence (CHSRA)			
facility admission profile			1.6 (1.5, 1.8)
Bowel & bladder incontinence high risk (CHSRA)			
facility admission profile			1.5 (1.4, 1.5)
Bowel & bladder incontinence low risk (CHSRA)			
facility admission profile			1.6 (1.5, 1.6)
Bowel incontinence change (LTCQ)			
short-term memory	57	2.1	1.6 (1.5, 1.8)
dressing problem	54	2.8	2.0 (1.9, 2.2)
bladder incontinence	43	2.7	1.8 (1.7, 2.0)
pressure ulcers	10	1.7	1.5 (1.4, 1.7)
facility admission profile			1.2 (1.1, 1.2)

Table 7.3.2
Association of Covariates and QI Prevalence at Resident Level, Massachusetts MDS Repository Data [n=554 facilities]

	Prevalence (per 100 res)	Crude OR	Adjusted OR[1] (95% CI)
QI			
Bladder incontinence change (LTCQ)			
short-term memory	57	1.8	1.6 (1.5, 1.7)
dressing problem	74	2.9	2.6 (2.3, 2.8)
decision-making problem	19	2.1	1.5 (1.3, 1.6)
weight loss	11	1.4	1.3 (1.1, 1.4)
facility admission profile			1.0 (1.0, 1.1)
Urinary tract infection prevalence (CHSRA)			
facility admission profile[a]			1.1 (1.1, 1.2)
Cognition			
Cognition change (LTCQ)			
eating self perform	17	1.7	1.4 (1.2, 1.7)
bowel incontinence	33	1.5	1.3 (1.1, 1.4)
fell last 30 days	19	1.3	1.3 (1.2, 1.5)
ability to understand others	08	1.9	1.4 (1.0, 1.8)
weight loss	13	1.4	1.2 (1.1, 1.4)
self initiated activities	59	1.3	1.1 (1.0, 1.3)
age greater than 76 years	73	1.3	1.2 (1.1, 1.3)
facility admission profile			1.1 (1.0, 1.1)
Communication			
Communication change (LTCQ)			
self perform eating	22	1.8	1.4 (1.2, 1.6)
short term memory	66	2.6	1.8 (1.6, 2.1)
self initiated activities	59	1.8	1.3 (1.2, 1.5)
motor agitation	12	1.5	1.2 (1.0, 1.4)
Alzheimer's disease	12	1.7	1.2 (1.1, 1.4)
wandering	11	1.6	1.1 (0.9, 1.2)
facility admission profile			1.2 (1.1, 1.3)
Drugs			
Antipsychotic prevalence (CHSRA)			
facility admission profile			1.7 (1.7, 1.8)
Antipsychotic prevalence high risk (CHSRA)			
facility admission profile			1.5 (1.4, 1.6)
Antipsychotic prevalence low risk (CHSRA)			
facility admission profile			1.7 (1.7, 1.8)
Anxiolytic/hypnotic prevalence (CHSRA)			
facility admission profile			1.5 (1.4, 1.6)
Falls			
Falls change (LTCQ)			
fell past 30 days	16	4.9	4.3 (4.0, 4.6)
fell past 31-180 days	21	2.5	2.0 (1.9, 2.2)

Table 7.3.2
Association of Covariates and QI Prevalence at Resident Level, Massachusetts MDS Repository Data [n=554 facilities]

	Prevalence (per 100 res)	Crude OR	Adjusted OR[1] (95% CI)
QI			
bedfast	83	4.3	2.9 (2.3, 3.8)
wandering	09	2.4	1.8 (1.6, 1.9)
facility admission profile			1.1 (1.0, 1.1)
Mobility			
Mobility Change (LTCQ)			
fell	37	1.6	1.4 (1.4, 1.5)
eating problem	22	0.4	0.4 (0.4, 0.5)
toileting problems	55	0.7	0.9 (0.8, 0.9)
bedfast	04	0.2	0.3 (0.2, 0.4)
facility admission profile			1.1 (1.1, 1.1)
Mood			
Depression without treatment (CHSRA)			
facility admission profile			1.5 (1.4, 1.6)
Depression with treatment (CHSRA)			
facility admission profile			1.4 (1.3, 1.5)
Depressed mood change (LTCQ)			
transfer function	68	1.3	1.1 (1.0, 1.2)
pain	48	1.1	1.2 (1.1, 1.2)
not a long-term resident	13	1.3	2.4 (2.1, 2.8)
mood not persistent	50	0.1	0.1 (0.0, 0.1)
depression	13	1.8	1.2 (1.1, 1.3)
facility admission profile			0.9 (0.9, 1.0)
Nutrition			
Tube feeding prevalence (Ramsey)			
ALS, CVA or Huntington's disease	17	3.4	2.3 (1.9, 2.7)
cognitive impairment	64	2.4	1.8 (1.4, 2.2)
swallowing problem	19	19.7	15.2 (12.4, 18.7)
Alzheimer's disease or other dementia	36	0.7	0.5 (0.4, 0.6)
facility admission profile			1.6 (1.4, 1.7)
Pain			
Pain change (LTCQ)			
decision making problem	44	1.4	1.3 (1.0, 1.7)
establishes own goals	13	1.5	1.4 (1.1, 1.7)
facility admission profile			1.1 (1.0, 1.3)
Pressure Ulcer			
Pressure ulcer prevalence (CHSRA)			
facility admission profile			1.3 (1.2, 1.3)
Pressure ulcer prevalence high risk (CHSRA)			
facility admission profile			1.2 (1.2, 1.3)

Table 7.3.2
Association of Covariates and QI Prevalence at Resident Level, Massachusetts MDS Repository Data [n=554 facilities]

	Prevalence (per 100 res)	Crude OR	Adjusted OR[1] (95% CI)
QI			
Pressure ulcer prevalence low risk (CHSRA)			
facility admission profile			1.3 (1.2, 1.4)
Pressure ulcer change (LTCQ)			
transfer problem	38	2.7	1.7 (1.5, 1.9)
unstable	23	1.4	1.4 (1.2, 1.5)
bed mobility problem	35	2.4	1.2 (1.1, 1.4)
locomotion problem	37	2.5	1.4 (1.2, 1.6)
bowel incontinence	39	1.9	1.1 (1.0, 1.2)
facility admission profile			1.3 (1.2, 1.3)
Restraints			
Restraint prevalence (CHSRA)			
facility admission profile			1.0 (1.0, 1.0)
Social			
Personal relationships change (LTCQ)			
short term memory problem	66	0.4	0.5 (0.3, 0.6)
bowel incontinence	5	2.7	1.8 (1.2, 2.7)
eating	7	3.3	1.9 (1.1, 3.3)
tearfulness	24	1.4	1.4 (1.0, 2.0)
verbally abusive	12	2.0	2.2 (1.5, 3.3)
facility admission profile			1.2 (1.0, 1.4)
Little or no activities (CHSRA)			
facility admission profile			2.2 (2.1, 2.2)
Weight			
Weight loss (LTCQ)			
long-term memory problem	52	1.1	0.9 (0.8, 1.0)
leaves 25% food uneaten	36	3.7	3.6 (3.2, 4.0)
bed mobility problem	40	1.6	1.3 (1.2, 1.5)
physically abusive	01	1.4	1.1 (0.8, 1.5)
facility admission profile			1.2 (1.2, 1.3)

res, residents; OR, Odds ratio; CI, Confidence Interval.

[1] Exponentiated logistic regression coefficient. Provides an index of the increase in likelihood (odds of satisfying QI definition) when specific dichotomous covariate is present or for a one standard deviation increase in facility admission profile covariate, holding constant the effect of the other covariates.

Prevalence and OR cannot be calculated for the facility admission profile variable since it is not a resident-level mean.

Table 7.3.3 provides evidence that the casemix adjustment models succeed in generating adjusted QIs that differ from the unadjusted (raw) QIs in ways that correspond to the level of resident acuity and facility measurement or selection effects. The first column of Table 7.3.3 lists the individual QIs. The second column provides the Pearson correlation coefficient describing the association of the raw (unadjusted) QI and the adjusted QI. Values near 1.0 suggest a perfect correlation between raw and adjusted QIs, and indicate the casemix adjustment procedure did little to alter the relative position of facilities on the QI. As values deviate from unity, the effect of casemix adjustment increases.

For example, the correlation of the raw and adjusted Tube Feeding QI scores is 0.49, which suggests that the adjusted value differs importantly from the raw value. Examination of columns three and four of the tables suggest this is largely due to facility-level differences in the proportion of residents admitted with tube feeding, rather than intrinsic resident-level risk factors. Notice that the association of the mean covariate score (the sum of the number of risk factors displayed by the resident, averaged over all residents in the facility) is almost zero (r=.08, no association) but the association with the admission profile variable (portion of residents among admissions with tube feeding) is quite high (r=0.80). Thus, nearly two-thirds (r^2=0.64) of the facility-level variability in tube feeding can be attributed to facility-level prevalence of tube feeding among admissions. Since there is not likely to be significant inter-facility variation in the ability of assessors to notice and record tube feeding, this effect can be seen as a selection bias effect. Facilities that admit a high proportion of residents that are tube fed will have high proportions of residents tube fed subsequently. Failure to account for facility variation in admission profile may lead to QI flags that are not driven by a problem of quality of care, but due to facility specific admission practices.[2]

[2] This approach does not address the issue of whether tubes should have been inserted in the hospital nor whether it would be preferable to have them removed after some period of time in the facility. However, these ethical issues are pertinent to the appropriateness of even having such a QI.

Table 7.3.3
Correlation of Unadjusted (raw) QI with Adjusted QI, Mean Facility Covariate Score, and Facility Admission Profile. Massachusetts Repository Data. [n=554 facilities]

	Correlation of unadjusted QI with		
	Adjusted QI	Mean Covariate Score	Admission Profile
OI			
ADL			
ADL decline (CHSRA)	1.00	na	.06
Behavior (CHSRA)	.86	na	.58
Behavior high risk (CHSRA)	.93	na	.48
Behavior low risk (CHSRA)	.80	na	.47
Behavior change (LTCQ)	.92	.33	.39
Catheter			
Indwelling urinary catheter (LTCQ)	.93	-.06	.39
Catheter (CHSRA)	.86	na	.45
Continence			
Bowel & bladder incontinence (CHSRA)	.93	na	.46
Bowel & bladder incontinence high risk (CHSRA)	.99	na	.32
Bowel & bladder incontinence low risk (CHSRA)	.86	na	.59
Bowel incontinence change (LTCQ)	.87	.34	.26
Bladder incontinence change (LTCQ)	.95	.25	.21
Urinary tract infection prevalence (CHSRA)	.99	na	.11
Cognition			
Cognition change (LTCQ)	.97	.27	.13
Communication			
Communication change (LTCQ)	.92	.35	.37
Drugs			
Antipsychotic prevalence (CHSRA)	.97	na	.60
Antipsychotic prevalence high risk (CHSRA)	.94	na	.30
Antipsychotic prevalence low risk (CHSRA)	.73	na	.64
Anxiolytic/hypnotic prevalence (CHSRA)	.98	na	.21
Falls			
Falls change (LTCQ)	.91	.45	.25
Mobility			
Mobility Change (LTCQ)	.99	.20	.09
Mood			
Depression without treatment (CHSRA)	.91	na	.29
Depression with treatment (CHSRA)	.95	na	.22
Depressed mood change (LTCQ)	.83	.37	.12
Nutrition			
Tube feeding prevalence (Ramsey)	.49	.05	.80
Pain			
Pain change (LTCQ)	.99	.10	.12
Pressure Ulcer			
Pressure ulcer prevalence (CHSRA)	.92	na	.39
Pressure ulcer prevalence high risk (CHSRA)	.97	na	.25
Pressure ulcer prevalence low risk (CHSRA)	.94	na	.26
Pressure ulcer change (LTCQ)	.91	.13	.33
Restraints			
Restraint prevalence (CHSRA)	1.00	na	.04
Social			
Personal relationships change (LTCQ)	.99	-.02	.05
Little or no activities (CHSRA)	.76	na	.72
Weight			
Weight loss (LTCQ)	.94	-.18	.23

NA, not applicable. CHSRA QIs, which have no resident-level risk covariates used in adjustment, have no corresponding facility mean covariate score.

Table 7.3.4
Mean, Distribution and Range of Facility Admission Profile Variables, Massachusetts Repository Data (n=554)

	Observed[1] distribution Mean (sd)	[min,max]
ADL Long Scale (Morris et al., 1999)	7.5 (1.8)	[0, 17]
Prevalence of verbally, physically abusive, or inappropriate behavior	0.2 (0.1)	[0, 1]
Sum of behavioral problems (E4a-e)	0.6 (0.6)	[0, 6]
Prevalence of indwelling catheter (H3d)	0.1 (0.1)	[0, 1]
Prevalence of either bowel or bladder incontinence (H1a>2 or H1b>2)	0.1 (0.1)	[0, 1]
Bowel continence (H1a)	0.3 (0.2)	[0, 1]
Bladder continence (H1b)	0.3 (0.2)	[0, 1]
Prevalence of urinary tract infection (I2j)	0.2 (0.1)	[0, 1]
CPS (Morris et al., 1994)	0.4 (0.2)	[0, 1]
Sum of items C4 and C6	0.4 (0.2)	[0, 1]
Prevalence of antipsychotic use O4a>0	2.0 (0.9)	[0, 6]
Prevalence of anxiolytic/hypnotic use	1.1 (0.7)	[0, 6]
Prevalence of falls	0.2 (0.2)	[0, 1]
Locomotion (G1ae)	0.3 (0.1)	[0, 1]
DRS (Burrows et al., 2000)	2.6 (0.6)	[0, 5]
Prevalence of tube feeding (L4b>0)	0.2 (0.1)	[0, 1]
Level of pain symptoms (J2b)	0.3 (0.2)	[0, 1]
Prevalence of pressure ulcer (M2a>0)	0.1 (0.1)	[0, 1]
Prev. restraint use (P4c-e equal to 2)	0.1 (0.1)	[0, 1]
Sum of personal relation items (F2a-d)	1.8 (0.3)	[1, 3]
Prevalence of no/little activities (N2=2|3)	0.4 (0.3)	[0, 1]
Weight loss (K3a>0)	0.2 (0.1)	[0, 1]

sd, standard deviation; min, minimum; max, maximum; CPS, Cognitive Performance Scale; DRS, Depression Rating Scale. Prev., prevalence
Item numbers refer to Resident Assessment Instrument/Minimum Data Set item numbers.
[1] observed minimum and maximum values in Massachusetts MDS repository data. Possible values often exceed observed.

7.4 Overview of QI Ranking by Steering Committee

The following section of this chapter describes the Steering Committee's final assessment of the value of the 26 evaluated QIs for different purposes: first, whether the measure would have utility for a nursing facility to assist in its quality improvement program; second, whether the QI would have utility and validity for use in guiding state surveyors as they go about the inspection process; and finally, whether consumers and purchases of service could use the measure to select facilities offering superior care services. Each rating was scored 0 to 7, with 0 signifying that the QI has very poor applicability and 7 meaning excellent applicability.

The tables present the Steering Committee rankings according to each QI's degree of acceptability for nursing facility continuous quality improvement (CQI) and long term care surveyor use and lists recommendations for future refinement of the QIs, including suggestions for dropping specific covariates. In applying the scores, the six voting Steering Committee (SC) members were guided by a consideration of the acceptability factors listed in Table 2.5. As use of the QIs moved out of the realm of an internal CQI to use in either the survey process or by consumers, the standards of acceptability become more strict.

The six sets of individual scores were averaged, and these average scores can be interpreted as follows:

$ A QI with an average score smaller than 3.0 has little support by the SC;
$ A QI with an average score greater or equal than 3.0 but smaller than 5.0 has moderate support; and
$ A QI with an average score greater or equal than 5.0 has strong support.

The Steering Committee also decided that any of the provisionally accepted 26 QIs that received a rating below 2.0 for the survey process should not be finally recommended to CMS, even if facilities might choose to utilize them for internal purposes. This reassessment of the original recommendation reflected the additional insight that was gained by applying the new risk-adjustment method to the existing QIs and affected four of those 26 QIs. The four QIs include Mood with Treatment (CHSRA), Mood without Treatment (CHSRA), Personal Relationships (LTCQ), and Antianxiety/ Hypnotic Use (CHSRA) and are described in the final section of this chapter, together with a brief explanation as to why the Steering Committee recommends them for use by CMS.

Table 7.4 presents a summary of the SC voting for the 26 QIs:

Table 7.4
Summary of Steering Committee Ratings on Audience Appropriateness for the 26 Approved QIs

Audience Use	Little Support (range 0.0-2.9)	Moderate Support (range 3.0-4.9)	Strong Support (range 5.0-7.0)
Nursing Facility CQI	0	10	16
Survey Process	4	12	10
Consumer/Purchaser	9	11	4

All 26 QIs received moderate to strong support for intra-facility CQI activities. Twenty-two received a similar level of support for use in the survey process, but the following five did not:

$ Prevalence of bladder and bowel incontinence (CHSRA),
$ Mood with treatment (CHSRA),
$ Mood without treatment (CHSRA),
$ Personal relationships (LTCQ), and
$ Antianxiety/hypnotic use (CHSRA).

As mentioned, the SC decided that the latter four QIs, even with the improved risk adjustment approach, could not be supported for any purpose but internal projects; thus they are not recommended for use by CMS.

The SC was most conservative in recommending QIs for use by consumers and purchasers of care. Unlike surveyor use, where there will be an on-site audit of the finding, family members or contracting agents working on behalf of the state or managed care companies will have no other recourse except to rely on the QI score itself. For this reason, the SC gave little support to nine of the 26 QIs, moderate support to 11 and strong support to the following four QIs:

$ Urinary tract infection (CHSRA),
$ Physical restraints (CHSRA),
$ Pressure ulcers (CHSRA), and
$ Indwelling catheter (CHSRA).

7.4.1 Functional QIs

Table 7.4.1 presents a summary of the Steering Committee ratings on audience appropriateness for the four functional QIs. For all QIs, the utility ratings for consumers were consistently lower than the corresponding scores for nursing facility internal quality assurance and survey use. The lower scores reflect fears of misclassifying a facility as having a problem due to inadequate casemix adjustment. With one the exception of the **Cognition** QI (LTCQ), all received moderate to strong support for use for all three audiences. **Locomotion** received the highest scores with 5.5 for nursing facility internal quality assurance, 4.6 for surveyors and 3.7 consumers.

These QIs were recommended for use but were not considered as being comprehensive measures for the concept of functional quality. For example, none of the QIs addressed residents' functional improvement. Facilities should be ranked by their ability to improve as well as prevent decline in residents' ADL function, locomotion, cognition, and communication. Future QIs for ADL could include the other ADL items such as dressing, bathing or the newer composite scales reported by Morris and colleagues (1999).

The high rating for locomotion was given despite the fact that the Steering Committee had several recommendations for future modifications and new QIs in this area. The limitation of the locomotion QI presented here is that independence in walking should be considered separately. Presently there is a danger that residents who change from one method of locomotion to another may be misclassified. Given that walking is considered by most to be preferable to wheelchair mobility, the change from walking to a wheelchair should be picked up by a QI. The present definition of locomotion would not necessarily detect such a change. Future locomotion QIs should separate walking and wheelchair independence as well as look to newer MDS 2.0 variables.

Table 7.4.1
Summary of Steering Committee Ratings on Audience Appropriateness for Functional QIs

QI	Nursing Facility	Survey	Consumer	Additional Comments/Recommendations
ADL Decline (CHSRA)	4.2	4.2	3	Future work will develop QI for improvement in ADL.
Cognition (LTCQ)	5.4	4.0	2.8	
Communication (LTCQ)	4.5	4.0	3.3	
Locomotion (LTCQ)	5.5	4.5	3.7	Recommend separating walking from wheelchair mobility as there is a possibility that residents may improve in walking ability yet appear more dependent on this QI because they require more help walking than wheeling their own wheelchair. Recommend development of a locomotion improvement QI.

7.4.2 Clinical QIs

Table 7.4.2 presents the Steering Committee's ratings for the 13 clinical complexity QIs with respect to their appropriateness for internal CQI in nursing facilities, use by the long-term care survey process, and use by consumers and purchasers of care. Overall, QIs in this domain received higher ratings for use in facility CQI vs. survey processes, and three of the 13 are not recommended for use by consumers. To illustrate, 10 of the 13 QIs were supported for the CQI process, but only nine for the survey process and four for use by consumers and purchasers. For the latter, the SC judged that caution is needed in determining whether a facility has a quality problem or conversely provides unquestioned superior care.

Both **Pressure Ulcer** QIs, representing prevalence **(CHSRA)** and change **(LTCQ)** measures, received high CQI ratings of 6.7 and 6.8, respectively, and both had survey ratings around 6 (6.7 and 6.0). Consumer ratings were only somewhat lower (5.3 and 4.5).

Likewise, for **Incontinence**, the two change (LTCQ) QIs received relatively high scores for both CQI (Bladder 5.8; Bowel 5.9) and the survey process (5.0), and a 4.2 (moderate support) rating for use by consumers. The prevalence **Bladder and Bowel Incontinence (CHSRA)** QI received lower scores in all categories (3.8, 2.7, 1.5). There was concern that, because the resident pool excludes residents having indwelling catheters, it could prompt facilities to increase use of catheters. Additionally, this QI identifies only residents with frequent or greater incontinence. Though risk-adjusted, the model may be underspecified.

The **Urinary Tract Infection (CHSRA)** QI had uniformly high scores assigned by the Steering Committee.

The **Falls (LTCQ)** QI was ranked as having moderate utility for CQI processes (4.8) and the survey process (3.7), but there was little support for its use as a consumer measure (1.5). The committee recommended exploring the use of MDS Version 2.0 measures of balance in revisions of the model.

The **Pain (LTCQ)** QI received a moderate rating for CQI (4.1) and surveyor use (3.7), but was not supported for use by consumers (2.2). This QI is limited in that it measures only new or worsening pain levels, leaving a gap for residents who are in persistent pain. There were also concerns about misclassification caused by under-recording of pain symptoms, particularly among persons who are unable to express their pain symptoms. Nonetheless, because this was the only pain QI forwarded for analysis it was recommended for use (albeit with caution, and then only for two of the three audiences) in order to raise the awareness to pain as a potential facility quality problem.

The **Feeding Tubes (Ramsey)** QI was given a rate of 5.0 for CQI and survey process use, and 4.8 for consumer use, reflecting the relevance of this topic.

Weight loss is a serious concern for nursing facility residents, and the **Weight Loss (LTCQ)** QI received a higher ranking for both CQI and the survey process (6.5), but not for public reporting (3.5). Some of the covariates may be better specified with MDS Version 2.0 data.

Both **Indwelling Catheter** QIs, representing prevalence **(CHSRA)** and change **(LTCQ)** measures, received high CQI ratings of 6.3 and 5.3, respectively. The CHSRA variant received universally high scores, while the LTCQ measure was judged to have only moderate support for use by surveyors and consumers. The concern with the LTCQ measure is that it is not stable over time and may be overadjusted. Nonetheless, chronic indwelling catheters can place individuals at risk of serious health complications. Recognizing this, the committee recommended both QIs for use.

Finally, the **Physical Restraints (CHSRA)** measure was viewed as being appropriate for use by nursing facilities, surveyors and consumers.

Table 7.4.2
Summary of Steering Committee Ratings on Audience Appropriateness for Clinical QIs

QI	Nursing Facility	Survey	Consumer	Additional Comments/Recommendations
Prevalence of Bladder and Bowel Incontinence (CHSRA)	3.8	2.7	1.5	Criteria for risk stratification may be better specified. From clinical and consumer points of view, identifying lesser levels of incontinence is also important for early intervention and could be considered for future work. Recommend dropping 2 service variable stratification criteria: bedrails and indwelling catheter.
Bladder Incontinence (LTCQ)	5.8	5	4.2	In view of tercile movement across adjacent time periods, the SC recommended the possibility of using "moving averages" vs. quarterly assessment of the QI. Because this QI only measures new/worsening incontinence, the SC may consider development of a QI that measures "improvement" for future work.
Bowel Incontinence (LTCQ)	5.9	5	4.2	In view of tercile movement across adjacent time periods, the SC recommended the possibility of using "moving averages" vs. quarterly assessment of the QI.
Urinary Tract Infection (CHSRA)	7.0	5.5	5.0	
Physical Restraints (CHSRA)	6.8	6.8	6.0	The SC believed that in view of the physical and psychosocial hazards associated with the use of physical restraints, survey teams should address the issue of restraints on an individual basis.
Feeding Tubes (Ramsey)	5.0	5.0	4.8	
Falls (LTCQ)	4.8	3.7	1.5	Recommend revising the model to include MDS Version 2.0 "balance" covariates.
Pain (LTCQ)	4.1	3.7	2.2	Develop a covariate-adjusted prevalence pain measure.
Stage 1-4 Pressure Ulcer (CHSRA)	6.7	6.7	5.3	Recommend refining the risk stratification with better specifications. Consider substituting the covariate, ICD-9 diagnosis of "malnutrition" with an MDS body mass index (BMI) score. There is likely ascertainment bias (underreporting) with the diagnostic measure; the BMI is a more accurate measure and would certainly help to identify more residents who are undernourished.
Pressure Ulcers (LTCQ)	6.8	6.0	4.5	
Indwelling Catheter (CHSRA)	6.3	6.3	6.3	
Indwelling Catheter (LTCQ)	5.3	4.0	3.2	
Weight Loss (LTCQ)	6.5	6.5	3.5	Recommend using "moving averages." If they are used, rather than quarterly estimates, the QI would be able to identify facilities with consistently poor quality in this area.

7.4.3 Psychosocial QIs

Table 7.4.3 presents the Steering Committee's rankings for the seven accepted psychosocial QIs. What stands out here is that two of the measures are not recommended for survey use and four are not recommended for use by consumers. The LTCQ **Mood** QI was rated as strongly appropriate for CQI (5.0) and moderately useful for the survey process and consumers. The two **Behavior** QIs (LTCQ and CHSRA) were supported for CQI and facility surveys, but there was little support for their use by consumers. The **Personal Relationships** (LTCQ) QI was ranked to have no utility for use by surveyors or consumers.

The **Mood** (CHSRA) QI was ranked as moderately useful for CQI (4.9) but not useful (0.2) in the survey process or for public reporting (0.0). Similarly, the **Mood with no treatment** (CHSRA) was ranked moderately (4.5) for use by nursing facilities and not useful for surveyors (0.0) or the public (0.0). The main concern was that the operational definition for both QIs identifies too few residents with depression so that *reported* rates are low and not indicative of the actually prevailing rates at a facility. Given the very restrictive definition, those two QIs are also prone to ascertainment bias, since only very astute assessors will document the necessary detail correctly. Finally, the **Little or No Activity** (CHSRA) QI had strong support for internal CQI, but little support as a measure for use by consumers**.**

Table 7.4.3
Summary of Steering Committee Ratings on Audience Appropriateness for Psychosocial QIs

QI	Nursing Facility	Survey	Consumer	Additional Comments/Recommendations
Behavior (CHSRA)	5.3	4.0	2.8	
Behavior (LTCQ)	4.5	3.7	2.3	
Mood (LTCQ)	5.0	4.7	3.0	
Mood (CHSRA)	4.9	0.2	0.0	Given its restrictive operational definition, this QI yields very low reported rates, which are unlikely to reflect the true rates of depression. Ascertainment bias was a big concern.
Mood with no treatment (CHSRA)	4.5	0.0	0.0	Given its restrictive operational definition, this QI yields very low reported rates, which are unlikely to reflect the true rates of depression. Ascertainment bias was a big concern.
Personal Relationships (LTCQ)	4.2	0.0	0.0	Future work will focus on alternative measurement to take into account previous levels of personal relationship difficulty and potential measurement other than annual.
Little or No Activity (CHSRA)	6.0	3.0	1.0	

7.4.4 Pharmacotherapy QIs

As displayed in Table 7.4.4, voting of Steering Committee members on the two recommended pharmacotherapy-related QIs was highly variable. **Use of Antipsychotic**

Drugs had moderate support for facility CQI and for use by consumers, but was not recommended for the survey process. Although the measure was adjusted for the facility admission profile, some SC members expressed concern about under-specification, as no explicit clinical information was incorporated in this QI. Additionally, this QI, in contrast to some of the CHSRA-developed QIs based upon MDS Section U data (detailed drug information), does not specify the type or the dose of the drug. Thus, applying this QI might unselectively discourage the use of antipsychotic drugs, which is why it was not recommended for the survey process.

SC members expressed greater concerns about the appropriateness of using the **Antianxiety/ Hypnotic** QI for aiding in the survey process with little support for using these measures by consumers and surveyors. While federal regulations and many state policies discourage use of certain classes of antianxiety/hypnotic agents, the scientific literature is actually far more circumspect in terms of condemning these drugs. Indeed, the cumulative evidence suggests that the greatest hazard is only related to use of high dose, long acting benzodiazapines, particularly as once a day doses in the evening. This is a far more restrictive definition of problematic use of this class of drugs, meaning that application of this QI in the survey process will detect many instances of *appropriate* uses of these drugs. Use in a facility-specific quality improvement process will be less subject to false interpretation and therefore was rated higher.

Table 7.4.4
Summary of Steering Committee Ratings on Audience Appropriateness for Pharmacotherapy QIs

QI	Nursing Facility	Survey	Consumer	Additional Comments/Recommendations
Anti-psychotic Use (CHSRA)	4.0	2.8	3.5	Future work will examine the relative importance of specific anti-psychotic agents by class and side effect profile, in order to determine the clinical validity of a general prohibition against all anti-psychotics, or whether some specific drugs are the most problematic.
Antianxiety/ hypnotic Use (CHSRA)	3.0	0.5	0.5	Future work will examine the relative prevalence of long acting, high dose prescriptions of these drugs since empirical research strongly suggests that it is this more limited range of uses and drug classes that can have the negative effects reported in the literature. We will be particularly focused on administration of single doses of long acting, high dose agents since use as sleeping aids apparently have the most pernicious effects.

7.4.5 QIs Rejected After Final Analyses

As mentioned above, the Steering Committee had decided not to finally recommend any QI that received a rating below 2.0 for the survey process. This final assessment took into consideration the results both from the initial conceptual and empirical review presented in

Chapter 6 and the results from the analyses incorporating the new risk-adjustment method presented in this chapter. The following four QIs were affected by this decision:

- $ Mood with treatment (CHSRA),
- $ Mood without treatment (CHSRA),
- $ Personal relationships (LTCQ), and
- $ Antianxiety/hypnotic use (CHSRA).

Both the **Mood with treatment** (CHSRA) and the **Mood without treatment** (CHSRA) QIs were primarily rejected because of their very low facility-level rates, when constructed from MDS V2.0 data. The Steering Committee agreed that those rates do not represent the true underlying rate of depression in nursing facilities and that variation in these QIs is therefore unlikely to reflect variation in quality of care. In addition, as the operational definition of these QIs is restrictive compared to DSM-IV criteria for depression, only very astute nursing staff will properly record all the items necessary to fulfill the QI definition. Thus, both QIs are prone to ascertainment bias.

The **Personal relationships** (LTCQ) QI was regarded as conceptually interesting but not sufficiently operationalized, since it does not take previous levels of social interaction into account. As it is constructed from annual data, the Steering Committee was concerned that it would not correctly identify important transient changes in the underlying concept. In addition, the measurement properties of this QI appeared questionable, since over 75 percent of the 554 facilities in the Massachusetts sample had a rate of zero and very few outliers had rates close to 1.0.

The **Antianxiety/hypnotic drug use** (CHSRA) QI was finally rejected on conceptual grounds. While the unselective use of sedative drugs, especially in the elderly, is associated with an increased risk of adverse event, such as falls or delirium, this QI cannot reliably distinguish between appropriate and inappropriate use of these agents. This QI generated much discussion among the project team and had been moved forward to this final analytic phase because it is one of the few drug-related QIs that can be constructed without MDS Section U data, which is not collected in all states.

8.0 Summary, Cautionary Notes about Implementation of the Recommended QIs, and Future Work

8.1 Summary of Findings

After an extensive review of both published and unpublished performance QIs and commercial quality measurement systems, a total of 44 QIs were selected for detailed testing and analysis. All 24 QIs incorporated by CMS into its new survey process (referred to throughout as the CHSRA QIs) were included in this evaluation. Several different databases covering six different states (Kansas, Maine, Massachusetts, New York, South Dakota, Vermont) were constructed in preparation for these analyses. Analyses conducted relied upon longitudinal files of MDS (either Versions 1.0, 1+ or 2.0) data which linked individual residents across multiple assessments within single facilities. The following issues were examined:

$ Prevalence or incidence of the QI among all nursing facility residents within and between states;
$ The actual number of residents at-risk of the QI by facility;
$ Distribution of the QI rates in all facilities in a state; and
$ The extent to which the relationship between the QI rate and the identified risk factors is comparable across states.

These analyses resulted in a total of 26 QIs provisionally recommended to CMS for generalized use in determining the quality of nursing facility care.

A second stage of analyses was undertaken to examine the effect that a number of different methodological complications might have on the potential meaningfulness and bias inherent in the QIs we recommended to CMS. The results of these supplemental analyses are documented in Chapter 7 and Technical Appendices 1 and 2. Emerging from these analyses was a unified approach to adjust for some of the major sources of bias inherent in applying QIs to facilities and comparing facilities' performance using QIs. In essence, by adjusting the facility QI for the proportion of residents admitted to a facility with the clinical problem addressed by the QI, facilities that specialize in such patients or that devote extra effort to identifying the problem are not penalized. This adjustment results in a QI that is less subject to the kinds of biases that serve to undermine the validity and therefore the utility of the QI.

Based on insight gained during this second stage of analyses, the project team reconsidered the 26 QIs that had initially been recommended and agreed that four of them no longer held up. Thus, the final number of recommended QIs is 22. Of these, nine were developed and are currently in use by CHSRA in various applications, twelve were developed and are currently in use by LTCQ and one was developed by another source (James Ramsey, Univ. Wisconsin). The recommended QIs are categorized as functional, clinical, psychosocial and pharmacotherapy QIs, and are described further in Chapter 6 as well as in Appendices 7 and 8. Table 8.1 lists all 22 QIs and their operational definitions.

Table 8.1
QI Definitions

Indicator	Numerator & Denominator Definition(s), and Exclusions	Covariate(s)	Developer
FUNCTIONAL QUALITY			
Late-Loss ADL worsening	**Numerator:** Residents with worsening (increasing item score) in ADL self-performance at current relative to previous assessment. Residents meet the definition of ADL worsening when at least two of the following are true: 1. G1a(A)[t]-G1a(A)[t-1] > 0, or 2. G1b(A)[t]-G1b(A)[t-1] > 0, or 3. G1h(A)[t]-G1h(A)[t-1] > 0, or 4. G1i(A)[t]-G1i(A)[t-1] > 0, or at least one of the following is true: 1. G1a(A)[t]-G1a(A)[t-1] > 1, or 2. G1b(A)[t]-G1b(A)[t-1] > 1, or 3. G1h(A)[t]-G1h(A)[t-1] > 1, or 4. G1i(A)[t]-G1i(A)[t-1] > 1. **Denominator:** Residents with valid contiguous assessments, excluding those totally dependent on ADL. (G1 a-j Box A - all ten items = 4 or 8) and also exclude comatose residents (B1=1). Note: for purposes of assessing change, ADL item (G1a,b,h,i box A) values of 8 are treated as missing data.	Facility admission profile: mean ADL Long Form (ADLLF) scale score among facility admissions over previous 9 months. <u>ADL Long Form</u> scale defined as sum of G1a(A), G1b(A), G1g(A), G1i(A), G1h(A), G1e(A), G1j(A), after converting 8's (did not do) to 4's (total dependence).	CHSRA
Cognition worsening	**Numerator:** Residents with score on cognitive performance scale (CPS, Morris et al. 1994, defined below) that is higher on current relative to prior assessment (CPS[t]>CPS[t-1]). **Denominator:** All residents with two contiguous assessments and valid data in required fields.	Bowel incontinence (H1a= 4) Fell past 30 days (J4a=checked) Weight loss (K3a=1) Age greater than 76 (see appendix on CALCAGE algorithm) Facility admission profile: mean CPS score among admissions over previous 9 months.	LTCQ

Table 8.1
QI Definitions

Indicator	Numerator & Denominator Definition(s), and Exclusions	Covariate(s)	Developer
	Cognitive Performance Scale: first define two interim/working variables, an "impairment symptom count" (ISC) and "severe impairment symptom count" (SISC). The ISC is the number of the following symptoms: 1. Short term memory problem (B2a=1) 2. Modified independence of moderately impaired daily decision making (B4=1 or 2), 3. Not always able to make self understood (C4>0). The SISC is the number of the following symptoms: 1. Moderately impaired daily decision making (B4=2), 2. Only sometimes or rarely makes self understood (C4>1). The CPS can then be defined as: 6 (very severe impairment) if comatose (B1=1) or Severely impaired in daily decision making (B4=3) and Totally dependent on others for eating (G1h(A)=4,8 or missing). 5 (severe impairment) if not CPS=6 and Severely impaired in daily decision making (B4=3) and not totally dependent on others for eating (G1h(A)<4). 4 (moderate/severe impairment) if CPS¹5 or 6 and SISC=2. 3 (moderate impairment) if CPS¹4,5 or 6 and ISC>1 and SISC=1. 2 (mild im 1 (borderline intact) if CPS is not 2 or higher and ISC=1. 0 (intact) if CPS is not 1 or higher and ISC=0.		
Worsening Communication	**Numerator:** Residents with a communication scale score (sum of 'ability to understand others' (C6) and 'making self understood' (C4)) that is greater at the current assessment relative to the previous assessment (C4[t]+C6[t] > C4[t-1]+C6[t-1]). **Denominator:** Residents with two valid and contiguous assessments for C4 and C6, and who do not have most impaired communication scale score at [t-1] (i.e., 6).	Eating problem extensive assistance or total dependence(G1h(A)= either 3,4,8) Short-term memory problem (B2a=1) Not at ease doing self-initiated activities (F1c=0) Facility admission profile: mean communication scale score among facility admissions over previous 9 months.	LTCQ

Table 8.1
QI Definitions

Indicator	**Numerator & Denominator Definition(s), and Exclusions**	**Covariate(s)**	**Developer**
Locomotion worsening	**Numerator:** Total number of residents whose value for locomotion self-performance is greater at current relative to previous assessment (G1e(A)[t]>G1e(A)[t-1]). **Denominator:** Residents with two valid and contiguous assessments for locomotion self-performance (G1e(A)), and who have a level of locomotion self-performance at previous assessment more independent than 'total dependence' or 'activity did not occur' (G1e(A)[t-1]<4).	Recent fall (J4a or J4b =checked) Extensive support or more dependence in eating (G1h(A)>2) Extensive support or more dependence in toileting (G1i(A)>2) Facility admission profile: mean level of mobility (G1e(A), treating 8's (activity did not occur) as 4's (total dependence)) among facility admissions over previous 9 months.	LTCQ
CLINICAL QUALITY			
Bladder or bowel incontinence prevalence	**Numerator:** Residents who were frequently incontinent or fully incontinent on most recent assessment (H1a=3 or 4, or H1b=3 or 4). **Exclusions, High & Low Risk:** admission and re-admission assessments, and also excluding residents who are comatose (B1=1), have indwelling catheter (H3d=checked), or have ostomy (H3i=checked). **Denominator, High risk:** Residents not otherwise excluded and with severe cognitive impairment (B4=3 & B2a=1), or totally dependent in mobility ADLs (G1aa=4,G1ba=4, & G1ea=4) at most recent assessment. **Denominator, Low risk:** All residents not otherwise excluded and not meeting the definition of High risk.	Facility admission profile: prevalence residents frequently or fully incontinent in either bowel or bladder (H1a=3 or 4, or H1b=3 or 4) among facility admissions over previous 9 months.	CHSRA
Worsening bladder continence	**Numerator:** Residents with a value for bladder incontinence greater at current assessment relative to previous assessment (H1b[t]>H1b[t-1]). **Denominator:** Residents with two valid assessments for H1b, excluding residents fully incontinent at previous assessment (H1b[t-1]=4).	Short-term memory problem (B2a=1) Dressing problem or did not occur (G1g(A) = either 3, 4, 8) Decision making problem (B4 = 3) Weight loss (K3a = 1) Facility admission profile: mean bladder incontinence (H1b) level among admissions over previous 9 months.	LTCQ

Table 8.1
QI Definitions

Indicator	Numerator & Denominator Definition(s), and Exclusions	Covariate(s)	Developer
Worsening bowel continence	**Numerator:** Residents with a value for bowel incontinence greater at current assessment relative to previous assessment (H1a[t]>H1a[t-1]). **Denominator:** Residents with two valid assessments for H1a, excluding residents fully incontinent at previous assessment (H1a[t-1]=4).	Short-term memory problem (B2a=1) Dressing problem or did not occur (G1g(A)=either 3, 4, or 8) Bladder incontinence (H1b=either 3 or 4) Pressure ulcers (M2a=either 1, 2, 3 or 4) Facility admission profile: mean bowel incontinence (H1a) level among admissions over previous 9 months	LTCQ
Prevalence of urinary tract infections	**Numerator:** Residents with urinary tract infection on most recent assessment (I2j = checked). **Denominator:** All residents on most recent assessment. **Exclusion:** Admission (Aa8a=1 or Aa8b=1) or significant change (Aa8a=3) assessments.	Facility admission profile: prevalence of urinary tract infection (i2j=checked) among admissions over previous 9 months	CHSRA
Restraints (physical) used daily, prevalence	**Numerator:** Residents who were physically restrained daily (P4c or P4d or P4e = 2) on most recent assessment. **Denominator:** All residents on most recent assessment.	Facility admission profile: prevalence of daily physical restraint use daily (P4c or P4d or P4e = 2) among admissions over previous 9 months.	CHSRA
Prevalence of feeding tubes	**Numerator:** All residents with a feeding tube at current assessment (K5b=checked). **Denominator:** All residents at current assessment.	Swallowing problem (K1b=checked) Multiple Sclerosis (I1w=checked) Alzheimer's disease or other dementia (I1q or I1u = checked) cognitive impairment (B4>0 or B2a=1) any of Amyotrophic Lateral Sclerosis (I3a-e=335.2), cerebrovascular accident (I1t=checked) or Huntington's disease (I3a-e=333.4) Facility admission profile: feeding tube prevalence (K5b=checked) among admissions over previous 9 months.	Ramsey

Table 8.1
QI Definitions

Indicator	Numerator & Denominator Definition(s), and Exclusions	Covariate(s)	Developer
Falls prevalence among those without recent history of falls	**Numerator:** Residents who had a fall in the last 30 days recorded on most recent assessment (J4a[t]=checked) **Denominator:** Residents who had assessments over 2 consecutive quarters, excluding those who had a fall recorded on their previous assessment (J4a[t-1]= checked).	Fell past 30 days (J4a=checked) Fell past 31-180 days (J4b=checked) Bedfast (G6a = 0) Wandering (E4a(A) = either 1,2,3) Facility admission profile: prevalence of falls (J4a) among admissions over previous 9 months.	LTCQ
Pain, worsening	**Numerator:** Residents with greater pain at current assessment relative to previous assessment, defined by greater score on pain scale. Pain scale defined as: 0 if J2a=0; 1 if J2a=1; 2 if J2a>1 & J2b=1, or 2; 3 if J2a>1 & J2b=3. **Denominator:** Residents with two valid assessments of pain intensity and frequency (J2a, J2b), and <u>excluding</u> residents with the highest level of pain at previous assessment (pain scale score=3).	Independent or Modified Independence in daily decision making (B4=0 or 1) Resident established their own goals (f1d=checked) Facility admission profile: mean pain scale score among admissions over previous 9 months (pain scale=0 if J2a=0; 1 if J2a=1; 2 if J2a>1 & J2b=1, or 2; 3 if J2a>1 & J2b=3).	LTCQ
Pressure ulcer (stage 1-4) prevalence	**Numerator:** Residents with pressure ulcers (Stage 1-4) on most recent assessment (M2a >0, or I3a-e = 707.0) **Denominator, High risk:** Residents that are impaired in bed mobility or transfer (G1a(A) or G1b(A)=3 or 4), comatose (B1=1), suffer malnutrition (I3a-e= 260, 262, 263.0, 263.1, 263.2, 263.8, or 263.9), or have end stage disease (J5c=checked) recorded on most recent assessment, excluding admission and re-admission assessments. **Denominator, Low risk:** All residents not satisfying inclusion criteria for High risk. **Exclusion:** Admission (Aa8a=1 or Aa8b=1) or significant change (Aa8a=3) assessments.	Facility admission profile: prevalence of stage1-4 pressure ulcers (M2a >0, or I3a-e = 707.0) among admissions occurring over previous 9-months.	CHSRA

Table 8.1
QI Definitions

Indicator	**Numerator & Denominator Definition(s), and Exclusions**	**Covariate(s)**	**Developer**
Worsening pressure ulcers	**Numerator**: Total number of residents whose value for M2a[t] is greater than value recorded on previous assessment (M2a[t-1]). **Denominator:** Residents with valid assessments of M2a at current and previous assessment. Residents with Stage 4 pressure ulcer at previous assessment are excluded (M2a[t-1]=4).	Transfer problem or did not occur (G1b(A) = 3, 4, or 8) Conditions/diseases make resident's functional status unstable (J5a = checked) Bowel incontinence (H1a = either 2,3, or 4) Bed mobility problem or did not occur (G1a(A) = either 3, 4, or 8) Locomotion problem or did not occur (G1e(A) = either 3, 4, or 8) Facility admission profile: prevalence of stage1-4 pressure ulcers (M2a >0, or I3a-e = 707.0) among admissions occurring over previous 9-months.	LTCQ
Prevalence of indwelling catheters	**Numerator:** Indwelling catheter on most recent assessment (H3d=checked). **Denominator:** Residents on most recent assessment, excluding admission & re-admission assessments.	Facility admission profile: prevalence of indwelling catheter among admissions over previous 9 months (H3d=checked)	CHSRA
New insertion of indwelling catheter	**Numerator:** Residents with an indwelling catheter (H3d[t]=checked) that did not have an indwelling catheter at previous assessment (H3d[t-1]= not checked). **Denominator:** Residents with two valid assessments of any of H3a-j (checked or not checked).	Bowel incontinence (H1a= 4) Pressure ulcers (M2a=either 1, 2, 3 or 4) Feeding tube (K5b=checked) Facility admission profile: prevalence of indwelling catheter among admissions over previous 9 months (H3d=checked)	LTCQ
Weight loss prevalence	**Numerator:** Residents with weight loss (K3a=1) on current assessment. **Denominator:** All residents at current assessment.	Long-term memory problem (B2b=1) Leaves 25% food uneaten (K4c=checked) Bed mobility problem or did not occur (G1a(A) = either 3, 4, or 8) Physically abusive (E4c(A) = either 2 or 3) Facility admission profile: prevalence of recent weight loss (K3a=1) among admissions over previous 9 months.	LTCQ

Table 8.1
QI Definitions

Indicator	**Numerator & Denominator Definition(s), and Exclusions**	**Covariate(s)**	**Developer**
PSYCHOSOCIAL QUALITY			
Behavior symptoms affecting others	**Numerator:** Residents with behavioral symptoms affecting others on most recent assessment, including any verbally abusive behavior (E4b(A)>0), physically abusive behavior (E4c(A)>0) or socially inappropriate behavior (E4d(A)>0). **Denominator, high risk:** Residents with cognitive impairment (B4>0 & B2a =1) or a psychotic disorder (I3a-e=295.00-295.9, 297.00-298.9, I1gg = checked), or bipolar disorder (I3a-e=296.00-296.9 or I1ff =checked) noted on the current or most recent full assessment. **Denominator, low risk:** Residents not otherwise excluded or satisfying the inclusion criteria for high risk. **Exclusion:** Admission (Aa8a=1 or Aa8b=1) or significant change (Aa8a=3) assessments.	Facility admission profile: mean of the sum of behavior item scores (E4a(A), E4b(A), E4c(A), E4d(A)) among facility admissions over previous 9 months.	CHSRA
Worsening behavioral symptoms	**Numerator:** Residents with more behavioral symptoms present at current assessment ([t]) relative to prior assessment ([t-1]). Included symptoms are Wandering (E4a(A)>0), Verbally abusive behavior (E4b(A)>0), Physically abusive behavior (E4c(A)>0), and Socially inappropriate behavior (E4d(A)>0). **Denominator:** Residents with two valid and contiguous assessments for indicated behavioral symptoms, excluding those with all four symptoms at previous assessment.	Modes of expression include speech (C3a=checked) Moderately or severely impaired cognitive skills (B4>1) Motor agitation (E1n>0) Facility admission profile: mean of the sum of behavior item scores (E4a(A), E4b(A), E4c(A), E4d(A)) among facility admissions over previous 9 months.	LTCQ

Table 8.1
QI Definitions

Indicator	Numerator & Denominator Definition(s), and Exclusions	Covariate(s)	Developer
Depressed/Anxious mood worsening	**Numerator:** The total number of residents whose Mood scale score is greater in current relative to previous assessment than previous assessment. Mood Scale: Sum of following eight conditions that are true: 1. Verbal expressions of distress (E1a>0, E1c>0, E1e>0, E1f>0, E1g>0, E1h>0), 2. Shows signs of crying, tearfulness (E1m>0), 3. Motor agitation (E1n>0), 4. Leaves food uneaten (K4c=checked), 5. Repetitive health complaints (E1h>0), 6. Repetitive/recurrent verbalizations (E1a>0, E1c>0, or E1g>0), 7. Negative statements (E1a>0, E1e>0, or E1f>0), 8. Mood symptoms not easily altered (E2=2). **Denominator:** Residents with two valid contiguous assessments for relevant items.	Transfer independent through extensive assistance (G1a(B)=0-3) Pain (J2a = 1, or 2) Discharge planned in 3 months (Q1c = 1 or 2) Facility admission profile: mean depression rating scale score (DRS; Burrows et al. 2001) among facility admissions over previous 9 months. Depression rating scale score defined as sum of E1a,d,f,h,i,l,m.	LTCQ
Little or no activity	**Numerator:** Residents with little or no activity (N2>1) on most recent assessment. **Denominator:** All residents (excluding comatose (B1=1) on most recent assessment. **Exclusion:** Admission (Aa8a=1 or Aa8b=1) or significant change (Aa8a=3) assessments.	Facility admission profile: prevalence of little or no activity (N2>1) among facility admissions over previous 9 months.	CHSRA

Table 8.1
QI Definitions

Indicator	**Numerator & Denominator Definition(s), and Exclusions**	**Covariate(s)**	**Developer**
PHARMACOTHERAPY QUALITY			
Prevalence of antipsychotic use in the absence of psychotic and related conditions	**Numerator:** Residents receiving antipsychotics (O4a>0) on most recent assessment. **Denominator, High risk:** All residents not otherwise excluded and suffering from both 'behavior problems' and 'cognitive impairment' (defined at right) **Denominator, Low risk:** Residents not excluded and not satisfying the definition for inclusion in 'high risk'. Exclusion: Admission (Aa8a=1 or Aa8b=1) or significant change (Aa8a=3) assessments, and residents with an indicated condition on the current, most recent, or most recent full assessment, including: 1. One or more psychotic disorders (I3a-e= 295.00-295.9, 297.00-298.9, I1gg=checked), 2. Tourette syndrome (I3a-e=307.23), 3. Huntington's (I3a-e=333.4) 4. Hallucinations are present (J1i = checked, current assessment only).	Facility admission profile: prevalence of antipsychotic use (O4a>0) among admissions over previous 9 months. *Definitions of denominator restriction terms (not used as covariates):* Cognitive impairment: Any impairment in daily decision making ability (B4 > 0) AND has Short-term memory problems (B2a = 1). Behavior problems : Defined as one or more of the following: 1. Verbally abusive behavior present (E4b(A) > 0), 2. Physically abusive behavior present (E4c(A) > 0), or 3. Socially inappropriate/disruptive behavior present (E4d(A)> 0).	CHSRA

In considering which QIs to recommend for use, the project Steering Committee gave precedence to QIs according to the following criteria:

- QIs that are applicable to a long-stay population;
- QIs with definitions that have face validity;
- QIs that have some form of risk adjustment;
- QIs that are applicable to a large segment of the nursing facility population; and
- QIs that address important clinical problems not addressed by any other recommended QI or QI currently in use by CMS.

As noted, our current recommendation is to apply a facility-level admission casemix adjuster to **all** of the QIs. A separate memorandum has been prepared for CMS which details the precise computer code and data file layout needed to construct these revised QIs, regardless of their original source.

As discussed in Chapter 2, there are numerous conceptual and statistical issues associated with the measurement of quality and the development of QIs. Inter-facility comparisons must consider casemix differences to address the selection bias that threatens the validity of the comparisons. Casemix differences can occur due to differential admission practices and to differential rates of discharge (censoring due to death, transfer or discharge home). Facilities may also be wrongly classified as performing "better" or "worse" due to assessment measurement errors. Such misclassification may result from facility differences in assessment skills, whereby better facilities detect more residents with a given problem and thus appear to be performing more poorly than facilities that have low rates because they failed to assess the problem. Facilities may also be misclassified if the QI rates used to assess their performance are unstable due to small numbers of residents included in the calculation of the QI. For example, facilities that are ranked within 30 percentile scores of each other may only have a difference of absolute rates of 0.05. To date, to our knowledge, no set of QIs in either the acute, ambulatory or the long-term care arena has been adequately tested to make sure that it fully addresses all of these issues.

We do recognize that the 22 recommended QIs, even if adjusted for the facility's admission profile, do not overcome all the methodological problems enumerated above. However, the Steering Committee believes that their application in the survey process and in some cases in public reporting adds considerably to regulators' and consumers' knowledge base. In spite of obvious limitations, it is the project team's opinion that using these adjusted QIs is better than not using them since they propel the industry forward while not unduly penalizing facilities for real differences in admission as well as assessment practices.

In the case of surveyors using the QI scores and rankings as a way to guide the long-term care survey, Steering Committee members were generally more positively inclined than they were in the case of public reporting, since surveyors would be able to see the facility and the records in a facility identified through QIs. This ability to "see behind the numbers" reassured some Steering Committee members that identifying a facility as a possible poor performer on the basis of the QIs could at least be checked by the surveyors' detailed visit to the home. There were questions about the applicability of quality rankings for an audience that would be making final decisions based upon data and scoring systems that were only valid, on average, rather than in each specific case. Thus, the Steering Committee was more concerned about the adequacy of the adjustment process and the validity of the resulting performance "ranks" for public reporting, precisely because consumers, their advocates and possibly purchasers would not give the home a "second chance", since they wouldn't visit a home with what might appear to be significant quality problems. Nonetheless, the project team believes that those QIs recommended for public reporting are sufficiently stable, meaningful and can be reasonably adjusted (on average) by the facility admission profile for public reporting to have validity, particularly when comparing the top and bottom performers.

The multi-dimensional nature of quality should also be kept in mind when reporting QIs to any audience.

8.2 Cautionary Notes Regarding the Use of Existing QIs in the Long Term Care Survey Process

Use of QIs in Survey Process. Current plans call for the use of MDS-based QIs in the survey process. While these data-based QIs represent a substantial improvement of the pre-survey information available to surveyors, caution should be exercised in their use. There are two primary types of problems associated with the use of data-based QIs: 1) surveyor understanding of measurement issues; and 2) the use of relative thresholds based on a peer group of other facilities within a state rather than standards of care as the benchmark for flagging a QI. The following discussion articulates the project team's concerns regarding how surveyors might interpret the QIs, and how the use of QIs to augment the quality of care survey as currently conceived may lead surveyors to target the wrong issues for investigation.

If the QIs are to be implemented as currently conceptualized and operationalized by CMS in its long-term care survey process, the QIs will identify not only facilities which exceed a set threshold for a particular QI (i.e., the 90th percentile), but the subset of nursing facility residents which triggered that QI. The goal of implementing the QIs in this way is twofold: 1) to focus the survey on quality of care (and eventually, quality of life) issues of particular relevance in a given facility; and 2) to provide a mechanism for the surveyor to select residents for quality of care review. Surveyors are instructed to select residents whose MDS data contributed to the calculation of the QI for which the facility was flagged. This method of resident selection presents several problems to both surveyors and long-term care facilities, including:

- the calculation of QI rankings does not occur in real-time; thus, the surveyor may target a resident or residents for inclusion in the quality of care review sample who no longer reside in the facility;
- the identification of individuals with particular QIs will tend to ascribe a label to those persons, creating the potential for them to be seen and reviewed only as "the QI", rather than as a complete individual with varying health and psychosocial needs; and
- care issues surrounding the nature of the QI, rather than the QI itself, are the more appropriate areas for investigation.

Rather than use QIs themselves to generate a resident sample, the project team would favor drawing a sample of residents with characteristics that place them "at risk" of having the indicated clinical condition. Thus, for pressure ulcers, selected residents might be those with a diagnosis of diabetes, those who are bed bound, incontinent or who had had a pressure ulcer in the past. Ideally, the sample should be based upon a "pull" of the MDS data in the facility on the day of the survey in order to minimize changes in residents' condition. Whether the cases are selected based upon the number of risks that they manifest or their specific conditions is an empirical issue. Research undertaken in 40 facilities that are part of the National HealthCorp chain over the last several years used this strategy to identify a sample at high risk of manifesting selected clinical problems that could, theoretically, be prevented in facilities that have strong preventative procedures in place. The researchers found that this was a very efficient way to draw a sample that allowed ample opportunity to observe how the facility provided care to residents with complex conditions.

Incomplete understanding by surveyors of data limitations, risk adjustment and stability of QIs could give them the impression that the facility percentile scores and other QI-based reports are more precise than is really warranted. QIs are suggestive, not exhaustive, and are subject to change. Surveyors must understand that the QI score pertains to the possibility of a problem area existing at the time the resident assessments were completed. At the time of the survey, these problem areas may or may not exist, but, more importantly, there may be new problem areas. Misperception about QI measurement could result in surveyors attributing too much weight to the pre-survey information and focusing on specific residents and/or problem areas to the exclusion of other, "new" problem areas.

The second problem relates to the use of relative thresholds as standards. This approach could produce a ceiling-like effect, by discouraging surveyors from examining areas in which the facility percentile scores were not "suspect." The measurement limitations described earlier can produce unintended results if too much reliance is place on the QIs. These QIs should not supersede standards of care mandated under OBRA's nursing home reform law.

QIs can greatly enhance the survey process if the surveyors understand their limitations and use them as one source of information to prepare for the inspection of nursing facilities. The survey process would be enhanced with the addition of surveyor training which includes a section on the meaning and limits of QIs in the survey process and guidance for using this information in conjunction with observational and other data collected during the long term care survey.

References

Albert M and Cohen C. 1992. “The test for severe impairment: An instrument for the assessment of patients with severe cognitive dysfunction.” J Am Geriatr Soc. 40:449-53.

Arling, G.; Karon, S. L.; Sainfort, F.; Zimmerman, D. R., and Ross, R. 1997. “Risk adjustment of nursing home QIs.” Gerontologist. Dec; 37(6):757-66; ISSN: 0016-9013.

Berlowitz, D. R.; Ash, A. S.; Brandeis, G. H.; Brand, H. K.; Halpern, J. L., and Moskowitz, M. A. 1996. “Rating long-term care facilities on pressure ulcer development: importance of case-mix adjustment [see comments].” Ann Intern Med. Mar 15; 124(6):557-63; ISSN: 0003-4819.

Bernabei, R.and Gambassi, G. 1998. “The SAGE database: introducing functional outcomes in geriatric pharmaco-epidemiology [letter].” J Am Geriatr Soc. 46(2), 251-2. ISSN: 0002-8614.

Bernabei, R.; Gambassi, G.; Lapane, K.; Landi, F.; Gatsonis, C.; Dunlop, R.; Lipsitz, L.; Steel, K., and Mor, V. 1998. “Management of pain in elderly patients with cancer.” SAGE Study Group. Systematic Assessment of Geriatric Drug Use via Epidemiology [see comments]. JAMA. Jun 17; 279(23):1877-82; ISSN: 0098-7484.

Bindman, AB. 1999. “Can Physician Profiles be Trusted?” JAMA. 281(22): 2142-2143.

Blumenthal, D. 1996. “Effects of Market reforms on Doctors and Their Patients.” Health Affairs Millwood. 15(2): 170-84.

Brandeis, G. H., Berlowitz, D. R., Hossain, M., and Morris, J.N. 1995. “Pressure Ulcers: The Minimum Data Set and the Resident Assessment Protocol.” Advances in Wound Care. 8(6):18-25.

Bravo, G. and Potvin, L. 1991. “Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions.” J Clin Epidemiol. 44(4-5):381-90; ISSN: 0895-4356.

Burrows, A.B., Morris, J.N, Simon, S.E., Hirdes, J. P., Phillips, C. 2000. “Development of an MDS-Based Depression Rating Scale for Use in Nursing Homes.” Age and Ageing.

Burrows, A. B., Morris, J. N., Simon, S. E., Hirdes, J. P., & Phillips, C. (2000). Development of a minimum data set-based depression rating scale for use in nursing homes. Age Ageing, 29(2), 165-72.

Castle, N.G., Mor, V., Banaszak-Hall, J. 1996. “Special Care Hospice Units in Nursing Homes.” The Hospice Journal. 12(3): 59-69.

Crooks, V. C.; Schnelle, J. F.; Ouslander, J. P., and McNees, M. P. 1995. “Use of the Minimum Data Set to rate incontinence severity.” J Am Geriatr Soc. Dec; 43(12):1363-9; ISSN: 0002-8614.

Davis, M.A. 1991. “On nursing home quality: a review and analysis.” Medical Care Review. 48:129-66.

Donabedian, A. 1966. “The Effectiveness of Quality Assurance.” Int Journal of Quality of Health Care. 8(4): 401-7.

Epstein, A. 1995. “Performance Reports on quality: prototypes, problems and prospects" NEJM. 333: 57-61.

Folstein, M.F., Folstein, S.E., McHugh, P.R. 1975. “‘Mini-Mental state’: A practical method for grading the cognitive state of patients for the clinician.” J Psychiatr Res. 12:1189-98.

Frederiksen, K.; Tariot, P., and De Jonghe, E. 1996. “Minimum Data Set Plus (MDS+) scores compared with scores from five rating scales.” J Am Geriatr Soc. Mar; 44(3):305-9; ISSN: 0002-8614.

Fries, B.E, and Galecki, A.1998. "The MDS Plausibility Index," in review for Health Services Research, November, 1998.

Fries, B.E., Simon, S.E., Morris, J.N.,, Flodstrum, C., and Bookstein, F.L. in press. “Pain in U.S. Nursing Homes: Validating a Pain Scale for the Minimum Data Set”. The Gerontologist.

Gambassi, G., Landi, F., Peng, L., Brostrup Jensen, C., Calore, K., Hiris, J., Lipsitz, L., Mor, V., & Bernabei, R. 1998. “Validity of diagnostic and drug data in standardized nursing home resident assessments: potential for geriatric pharmacoepidemiology.” Med Care. 1998 Feb; 36(2):167-79; ISSN: 0025-7079.

Hartmaier, S. L., Sloane, P. D., Guess, H. A., and Koch, G. G. 1994. “The MDS Cognition Scale: a valid instrument for identifying and staging nursing home residents with dementia using the minimum data set [see comments].” J Am Geriatr Soc. 42(11):1173-9.

Harrington, C., Carrillo, H., Thollaug, S.C., Summers, P.R. 1999. Nursing Facilities, Staffing, Residents, and Facility Deficiencies, 1991 through 1997. Report produced by the Department of Social and Behavioral Sciences of the University of California at San Francisco.

Harrington, C., Zimmerman, D., Karon, S., Robinson, J., and Beutel, P. 1999. Nursing Home Staffing and its Relationship to Deficiencies. Report produced by the Department of Social & Behavioral Sciences of the University of California at San Francisco, and the Center for Health Systems Research and Analysis of the University of Wisconsin-Madison.

Hawes, C.; Morris, J. N.; Phillips, C. D.; Fries, B. E.; Murphy, K., and Mor, V. 1997. "Development of the nursing home Resident Assessment Instrument in the USA." Age Ageing. Sept.; 26 Suppl 219-25; ISSN: 0002-0729.

Hawes, C.; Morris, J. N.; Phillips, C. D.; Mor, V.; Fries, B. E., and Nonemaker, S. 1995. "Reliability estimates for the Minimum Data Set for nursing home resident assessment and care screening (MDS)." Gerontologist. Apr; 35(2):172-8; ISSN: 0016-9013.

Hirth, Banaszak-Holl and MCCarthy. 1999. "Nursing Home to Nursing Home Transfers: Prevalence, Timing and Correlates." Paper presented at 16th Annual Meeting of Association of Health Services Research, June 28, Chicago, Illinois.

Hofer, T. P., Hayward, R. A., Greenfield, S., Wagner, E. H., Kaplan, S. H., & Manning, W. G. 1999. "The unreliability of individual physician 'report cards' for assessing the costs and quality of care of a chronic disease." JAMA. 281(22): 2098-2105.

Institute of Medicine (IOM). 1990. "Effectiveness and outcomes in health care." K.A. Heithoff and K.N. Lohr, editors. National Academic Press, Washington, D.C.

Intrator, O.; Castle, N. G., and Mor, V. 1999. "Facility characteristics associated with hospitalization of nursing home residents: results of a national study." Med Care. Mar; 37(3):228-37; ISSN: 0025-7079.

Kahn, H., & Sempos, C. (1989). Statistical methods in epidemiology. New York: Oxford University Press.

Kiel, D. P.; O'Sullivan, P.; Teno, J. M., and Mor, V. 1991. "Health care utilization and functional status in the aged following a fall." Med Care. Mar; 29(3):221-8; ISSN: 0025-7079.

Kramer, A. 1999. Manual of Operations for "Staged Quality of Care Survey." University of Colorado Health Sciences Center.

Landi, F.; Gambassi, G.; Lapane, K. L.; Sgadari, A.; Gifford, D.; Mor, V., and Bernabei, R. 1998. "Comorbidity and drug use in cognitively impaired elderly living in long-term care." Dement Geriatr Cogn Disord. Nov-Dec 31; 9(6):347-56; ISSN: 1420-8008.

Landi, F., Gambassi, G., Lapane, K. L., Sgadari, A., Mor, V., & Bernabei, R. 1999. "Impact of the type and severity of dementia on hospitalization and survival of the elderly." Dement Geriatr Cogn Disord. 10: 121-9.

Lipowski, E. E. and Bigelow, W. E. 1996. "Data linkages for research on outcomes of long-term care." Gerontologist. Aug; 36(4):441-7; ISSN: 0016-9013.

Maxwell, C., Zimmerman, D., Karon, S., and Sainfort, F. 1998. "Estimating QIs for chronic care from the MDS 2.0: risk adjustment issues and concerns." Can J Quality Health Care. 14(3): 4-13.

Miller, S. C., Gozalo P., and Mor V. 2000. "Outcomes and Utilization for Hospice and Non-Hospice Nursing Facility Decedents." Report to the U.S. Department of Health and Human Services (HHS), Office of Disability, Aging and Long-Term Care Policy (DALTCP). Contract #100-97-0010.

Mor, V. 1999. "Hospitalization of nursing home residents: a review of clinical, organizational and policy determinants." Brown University, Center for Gerontology and Health Care Research Internet Site.

Mor, V., Branco, K., Fleishman, J., Hawes, C., Phillips, C., Morris, J., & Fries, B. (1995). The structure of social engagement among nursing home residents. Journals of Gerontology B: Psychological Sciences and Social Sciences, 50(1), 1-P8.

Mor, V.; Intrator, O.; Fries, B. E.; Phillips, C.; Teno, J.; Hiris, J.; Hawes, C., and Morris, J. 1997. "Changes in hospitalization associated with introducing the Resident Assessment Instrument [see comments]." J Am Geriatr Soc. Aug; 45(8):1002-10; ISSN: 0002-8614.

Mor, V.; Intrator, O., and Laliberte, L. 1993. "Factors affecting conversion rates to Medicaid among new admissions to nursing homes." Health Serv Res. Apr; 28(1):1-25; ISSN: 0017-9124.

Mor, V., Morris, J., Lipsitz, L., and Fogel, B. 1997. "The Q-Metrics system for long-term care outcomes management." Nutrition. 13(3): 242-4.

Mor, V., Banaszak-Holl, J., and Zinn, J.S. 1996. "The trend toward specialization in nursing care facilities." Generations, Winter 1995-96: 24-9.

Morris, J. N.; Fries, B. E.; Mehr, D. R.; Hawes, C.; Phillips, C.; Mor, V., and Lipsitz, L. A. 1994. "MDS Cognitive Performance Scale." J Gerontol. Jul; 49(4):M174-82; ISSN: 0022-1422.

Morris, John N., Fries, B. E., Morris, S. A. Scaling ADL's Within the MDS. Journals of Gerontology: Medical Sciences, 54A(10);1999.

Morris, J. N.; Hawes, C.; Fries, B. E.; Phillips, C. D.; Mor, V.; Katz, S.; Murphy, K.; Drugovich, M. L., and Friedlob, A. S. 1990. "Designing the national resident assessment instrument for nursing homes." Gerontologist. Jun; 30(3):293-307; ISSN: 0016-9013.

Morris, J. N.; Nonemaker, S.; Murphy, K.; Hawes, C.; Fries, B. E.; Mor, V., and Phillips, C. 1997. "A commitment to change: revision of CMS's RAI [see comments]." J Am Geriatr Soc. Aug; 45(8):1011-6; ISSN: 0002-8614.

Morris, J. N. 1999. Personal communication.

Morris, J.N., Fries, B.E., Morris, S.A. In Press. "Scaling ADLs within the MDS." Journal of Gerontology: Medical Sciences.

Morris, J.N., Fiatarone, M., Kiely, D.K., Belleville-Taylor, P., Murphy, K., Littlehale, S.. Ooi, W.L., O'Neill, E., and Doyle, N. In press. "Nursing Rehabilitation and Exercise Strategies in the Nursing Home." Journal of Gerontology: Medical Sciences.

Mukamel, D. B. and Brower, C. A. 1998. "The influence of risk adjustment methods on conclusions about quality of care in nursing homes based on outcome measures." Gerontologist. Dec; 38(6):695-703; ISSN: 0016-9013.

Mukamel, D.B. and Spector, W.D. 1998. "Nursing home costs and risk adjusted outcome measures of quality." Presented at the 51st Annual Scientific Meeting of the Gerontological Society of America. Philadelphia, PA.

Nyman, J. A. 1988. "Improving the quality of nursing home outcomes. Are adequacy- or incentive-oriented policies more effective?" Med Care. Dec; 26(12):1158-71; ISSN: 0025-7079.

Ooi, W. L., Morris, J. N., Brandeis, G. H., Hossain, M., & Lipsitz, L. A. 1999. "Nursing home characteristics and the development of pressure sores and disruptive behaviour." Age Ageing. 28: 45-52.

Phillips, C. D.; Chu, C. W.; Morris, J. N., and Hawes, C. 1993. "Effects of cognitive impairment on the reliability of geriatric assessments in nursing homes." J Am Geriatr Soc. Feb; 41(2):136-42; ISSN: 0002-8614.

Phillips, C. D.; Morris, J. N.; Hawes, C.; Fries, B. E.; Mor, V.; Nennstiel, M., and Iannacchione, V. 1997. "Association of the Resident Assessment Instrument (RAI) with changes in function, cognition, and psychosocial status [see comments]." J Am Geriatr Soc. Aug; 45(8):986-93; ISSN: 0002-8614.

Phillips, C. D.; Zimmerman, D.; Bernabei, R., and Jonsson, P. V. 1997. "Using the Resident Assessment Instrument for quality enhancement in nursing homes." Age Ageing. Sep; 26 Suppl 2: 77-81; ISSN: 0002-0729.

Porell, F. and Caro, F. G. 1998. "Facility-level outcome performance measures for nursing homes." Gerontologist. Dec; 38(6):665-83; ISSN: 0016-9013.

Shapiro, E., & Tate, R. B. 1995. "Monitoring the Outcomes of Quality of Care in Nursing Homes Using Administrative Data." Canadian Journal on Aging. 14(4): 755-768.

Spector, W. D., & Mukamel, D. B. 1998. "Using outcomes to make references about nursing home quality." Evaluation & the Health Professions. 21(3): 291-315.

Swanson, E. A. and Glick, O. J. 1995. "Reliability and validity of a new preadmission acuity tool for long-term care." J Nurs Meas. Summer; 3(1):77-88; ISSN: 1061-3749.

Zimmerman, D. 1999. Presentation at the Meeting on QI Validation, Hebrew Rehabilitation Center for the Aged, 11 March.

Zimmerman, D., Karon, S., Swearengen, J. 1999. "The Quality Component of the Demonstration," in DRAFT: CMS NHCMQ Demonstration Final Report, March.

Zimmerman, D. and Karon, S. 1998. Telephone interview with Terry Moore, Abt Associates Inc., 26 January.

Zimmerman, D. and Karon, S. 1997. "Measuring the Quality of Nursing Home Care: Risk Adjustment, Validation, and other 'Trivial' Issues." Presentation to the National Case Mix Reimbursement and Quality Assurance Conference, 22 September.

Zimmerman, D., Karon, S., Arling, G. et al. 1995. "Development and Testing of Nursing Home QIs." Health Care Financing Review. 19(4):107-127.

Zinn, J. S. 1994. "Market competition and the quality of nursing home care." J Health Politics Policy Law. Fall; 19(3):555-82; ISSN: 0361-6878.

Zinn, J. S.; Aaronson, W. E., and Rosko, M. D. 1993. "The use of standardized QIs as quality improvement tools: an application in Pennsylvania nursing homes." Am J Med Qual. Summer; 8(2):72-8; ISSN: 1062-8606.

Zinn, J. S.; Aaronson, W. E., and Rosko, M. D., 1993. "Variations in the outcomes of care provided in Pennsylvania nursing homes. Facility and environmental correlates." Med Care. Jun; 31(6):475-87; ISSN: 0025-7079.

Technical Appendix A

Supplemental Methodological Analyses

This appendix describes in detail analyses that were undertaken to examine technical issues that are related to the use of QIs. The issues are relevant regardless of the QI structure or the specific elements used to define the QI or the populations to which the QI applies. These issues, raised conceptually in the second chapter of this report, were examined empirically using available data from several different sources. In some cases the issues raised will necessarily call into question the validity of any QI. However, even a QI that can not overcome all the conceptual and technical issues outlined below may still have value and validity if properly applied.
Topics covered in this chapter include:

$ casemix differences;

$ variation in assessment skills or recording practices that can result in ascertainment bias;

$ effects of the size of the available resident pool from which to derive a QI and the stability of that estimate over time, as well as the associated error in establishing facility ranking estimates of the QI rates; and

$ the lack of homogeneity in the casemix profile across facilities.

A.1 Casemix Differences: Evaluating the Presence of Selection or Attrition Bias

Inter-facility comparisons of quality should be fair and unbiased, and we must be alert to the possibility that factors other than the care being provided influence our estimates of quality. Since most of the QIs reflect rates of adverse events, a likely factor to influence QI rates is the average health and functional status of residents in a facility, or the facility's casemix. Differences in casemix can be introduced through a variety of mechanisms: specialized admission practices (which can result in selection bias); differences in facility exit rates (which can result in attrition bias); and simple over-time variations in the casemix distributions of factors that relate to the likelihood that residents will acquire the problems tracked by the QIs. The following sections address these issues.

A.1.1 Differential Admission Practices and the Risk of Selection Bias

Facilities should not be reprimanded for potential problematic outcomes that were inherited at admission, nor should they be praised because of positive attributes that derive from the targeted admission practices of the facility. Differences in admission characteristics of residents may result from the propensity of certain facilities to admit very complex or specialized types of residents, while other facilities admit lighter care, less complicated persons into residency. The differences in admission characteristics may be due to referral patterns from regional hospitals, investments made by the nursing facility to provide special services or to variation in the availability of other local providers such as home care agencies.

Marked differences in referral patterns and admission practices could be especially problematic, as existing QIs have not been designed to account for them. Consequently, assessment of facility quality based on such unadjusted QIs may be biased against facilities who admit sicker residents.

A.1.2 Test for Differential Admission Practices

Measures of resident status at admission are parameters for which the nursing home should not be held accountable. It is unreasonable to expect facilities to have affected such measures during the first few days of residency, and it is on this basis that the CHSRA QIs exclude admission assessments. But, while this is an appropriate step, is it enough; or, does the lingering affect of differential admission practices continue to contaminate the QI estimate for the facility?

To address this question, analyses were undertaken to determine whether facility admission practices bias subsequent cross-sectional estimates of facility performance. Using data from over 500 nursing homes in the state of Massachusetts, we first derived two independent summary measures of resident performance for each facility.

$ The first is based on an aggregate measure derived from the intake MDS assessments for all admissions over a nine-month period. For example, for the measure weight loss over the prior 30 days, we would go to the MDS 2.0, Section K, item 3a, “Weight Loss.” For all admission assessments we would add one to the denominator, and if the resident was scored as “1" (recent weight loss) we would add one to the numerator. Thus, if there were 40 admission assessments over the nine month period, and 12 of these residents had a score of “1" on weight loss for the period immediately prior to the admission assessment, the estimate of the facility’s selection proclivity for these types of residents would be 12/40, or a proportional score of .30.

$ The second facility summary measure was derived from the cross-sectional cohort of all residents who had been in the facility for a minimum of nine months. With this rule, there is no overlap in the samples used to derive the two facility estimates.

A diverse set of measures were selected for this test, including: ADL performance (using the ADL Long Form, Morris, Fries, Morris, 1999); Cognitive performance (using the CPS, Morris, et al, 1994); Pain; Falls; Weight Loss; Use of Trunk Restraints; and Pressure Ulcers. For each measure, differential admission practices were considered to be a potential source of bias if there was a reasonably high correlation between the two facility’s aggregates for each measure -- intake and cross-sectional. The following table (Table A.1.1) displays these correlations; Figure A.1.1 shows a graphic view of these relationships for two of measures.

Table A.1.1

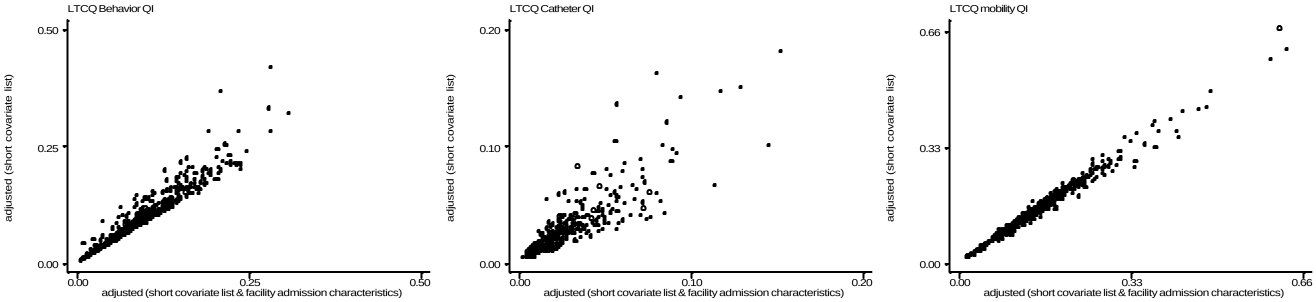
Measure	Correlation at Facility Level between the Estimate for the Measure at Intake and for a Cross-Sectional Cohort [n=554 MA facilities]
ADL Long Form	.61
Cognitive Performance Scale	.61
Pain	.59
Weight Loss	.24
Falls	.30
Use of Trunk Restraints	.53
Pressure Ulcers	.35

The correlation coefficients values ranged from a low of .24 to a high of .61, or from a moderate to high level. The value was highest for the functional parameters (ADLs and Cognition) and pain. But it was also a factor for the resolvable clinical complications of pressure ulcers and falls, as well as for the process of care measure of restraint use. Thus, differences in casemix are not adequately accounted for by excluding measures derived from the initial intake assessment. Rather, a more complex adjustment procedure is needed to address the residual effects of selection bias. While some of the existing QI measures we reviewed in this report included patient-based covariates, none of the measures specifically adjust for the selection bias identified above.

Including facility-level admission acuity measures in the resident risk adjustment models offers an additional layer of casemix adjustment. It accounts for differential admission practices that can lead to biased QI measures, as facilities which admit residents of high acuity will tend to have high QI values, even if they provide excellent care. It may also correct for differences in coding accuracy that can lead to ascertainment bias, as facilities with skilled assessment staff detect a higher share of the existing problems and thus appear to have higher QI values than their less diligent counterparts. The results of this additional level of casemix adjustment is summarized in Figure A.1.1, below.

Figure A.1.1

Scatterplots for three adjusted QIs; short covariate list adjusted (y-axes) and short covariate list and admission characteristic adjusted (x-axes) QIs, Massachusetts repository data.



A)LTCQ Behavior (r=0.96)

B) LTCQ Catheter (r=0.81)

B) LTCQ Mobility (r=0.99)

The plots in Figure A.1.1 display the QI values for facilities in the Massachusetts repository. Along the y-axis are QI values adjusted for a shorter list of covariates in keeping with the strategy detailed in Chapter 5. Along the x-axis are QI values adjusted not only for the shorter list of covariates but also a facility-level admission acuity covariate. Aggregates of facility-level data are included in the resident-level risk adjustment models, thereby adjusting for level of acuity of the facility's residents at admission, and also in part for the ability of the staff to detect and record subtle signs and symptoms among their residents. For the three QIs shown in Figure A.1.1, the means and standard deviations for the two methods of adjustment are very similar. The adjusted and adjusted-with-facility-effect QIs are highly correlated. In panel A, the Behavior QI, we notice that some facilities with apparently high adjusted rates have rates closer to the mean when facility effects are taken into consideration. In panel B, catheters, we see larger differences between adjusted and adjusted-with-facility-effect QIs, and hypothesize that this difference is due to differences in resident acuity at admission. For Mobility, notice that the two methods of adjustment produce nearly identical QI values. This plot emphasizes the point that not all QIs are equally susceptible to the residual effects of selection bias.

A.1.3 Differential Censoring (death, discharge and transfer)

Conceptually, facilities should be accountable for all patients for whom they care. If some facilities differentially discharge patients who are similar to those who are retained by other facilities, QIs that do not take those differences into account might be biased. The key issue is whether the discharges (or deaths) were related to the quality of care provided by the discharging facility. Residents may die, be re-hospitalized or transferred to other long term care facilities because of poorer quality of care such as inadequate medical management or failure to intervene appropriately. Facilities with high exit rates due to inadequate medical management would be difficult to assess with any sense of confidence. With such discharge practices, one would be hard pressed to hold the facility responsible for providing poorer care, as existing QIs make no provision for patients who die or are discharged during the measurement period. Indeed, QI rates for these problematic facilities would appear lower than for facilities who continue to care for sicker or heavy care residents.

Recently-published data reveal that facilities vary substantially in terms of the rate of hospitalization of their residents. Facilities that have more skilled nursing and medical staff have lower hospitalization rates (Intrator et al, 1999). Re-analysis of the 10-state MDS data set collected in the early 1990s for CMS's initial roll out of the MDS revealed large inter-state variation in hospitalization rates among nursing facility residents in the last six months of their lives (Mor, 1999).   Furthermore, analyses of the role of hospice care in the nursing facility revealed that nursing facilities which have a relationship with hospice care have far lower hospitalization rates than do facilities without such a relationship (Miller, Gozalo and Mor, 2000). While inter-facility transfers are relatively rare (less than two percent of all transfers from the nursing facility), since they may be concentrated in a minority of facilities, the observed differences in hospitalization rates and even mortality rates clearly could bias the measurement of any QI predicated on only the resident population (Mor, et al 1997).

The project team was concerned about the implications of these studies for the construction and application of valid QIs. If some facilities discharge residents who begin to manifest signs of negative outcomes, the true extent of quality problems in these facilities will be underestimated. Even risk adjustment (as described elsewhere in this report) is unable to control for this phenomenon since discharged, or censored, residents will not be included in the quality measurement.

The project team used two different approaches to investigate the potential impact of censoring bias. First, we constructed a data file using the Nursing Home Casemix and Quality (NHCMQ) demonstration MDS+ data files merged with CMS denominator and claims files. All residents of all nursing facilities in five states (Kansas, New York, Maine, Mississippi and South Dakota) who had an MDS assessment near January 1, 1996 were followed for up to six months to determine whether and when they had a subsequent MDS assessment. All intervening hospitalizations, deaths or uses of other Medicare benefits that are not covered when delivered in the nursing facility (e.g., home health) were noted. This made it possible to construct a "follow-up" measure that combined subsequent quarterly MDS records with data on death and hospitalization. This made it possible to determine the reason for censoring of cases that did not have a completed MDS. The project team separately tested the effect of censoring on new admission and long-stay residents (those with a stay greater than 90 days).

The project team also created a facility-level file based upon all the resident assessments in the five states during 1996. For selected QI-defined "outcomes" of interest (e.g., risk-adjusted worsening daily pain), the proportion of residents in the home with the event was correlated to the proportion of residents starting the baseline period who were no longer in the facility 12 months later. We used the long-stay population, since it is clear from many published sources and our own analyses of these data that the admission rate of nursing facilities is strongly related to the discharge rate. We didn't differentiate between death and hospitalization because states outside of the five NHCMQ demonstration states had not been matched to CMS claims.

In addition, the project team examined the relationship between the presence of cancer or diabetes and pain or pressure ulcers in order to understand whether the factors which place a resident at risk of incident pain or pressure ulcer also place them at risk of death, rehospitalization or other discharge. State differences in censoring were also examined. Finally, a logistic regression analysis was undertaken to test for the resident factors associated with censoring among residents without pressure ulcers, who were at risk of acquiring them.

The proportion of residents in daily pain was correlated with the proportion of residents still in the facility 12 months after baseline. Since we observed a significant, although not large, correlation between these two measures, we performed an OLS regression to control for other factors (e.g., acuity) to determine whether the relationship between outcome and censoring persists once other casemix factors had been considered. The results of these analyses are presented below.

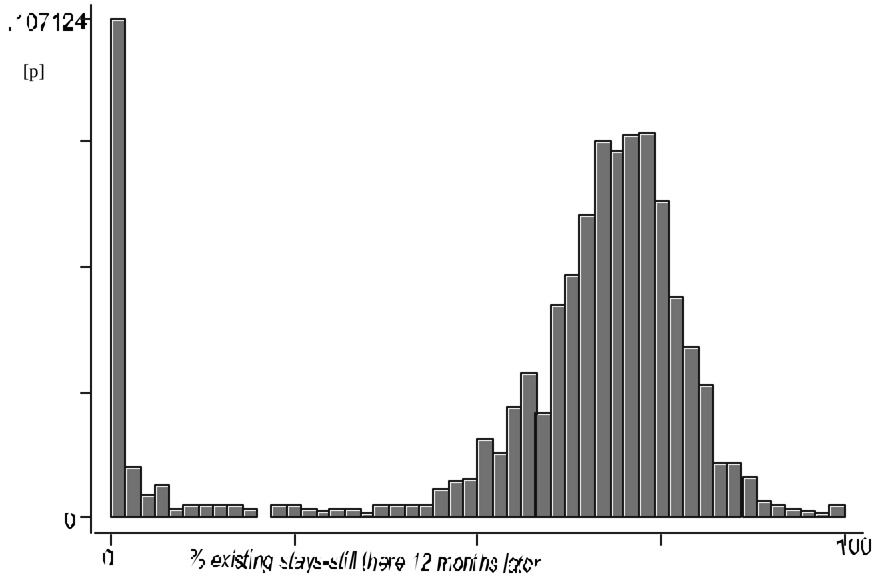
Only a small number of long stay patients were lost to followup during any particular assessment interval. Nearly 90 percent of the over 62,000 observations in the database had a 90-day follow-up assessment within 60 days from the "baseline" assessment. Just over four percent of residents died prior to their scheduled assessment. Relatively few residents missed an assessment due to being hospitalized (1.55 percent) and virtually none returned home to use the home health benefit. Interestingly, over three percent of residents did not have an assessment within the designated time frame, suggesting that departures from the required periodicity of assessments might play a role in the interpretation and meaningfulness of QIs.

The median facility in the file of over 1,500 facilities had 67 percent of its residents still in the facility 12 months after the baseline. At the 75th percentile of this distribution of facilities, 73 percent of residents remained in facilities at 12 months. Figure A.1.2 summarizes the one-year loss distribution among the 1,500 facilities. As is apparent, some facilities have very high rates of loss to follow-up. These facilities will have few observations to contribute to the creation of QI scores. This graph does reinforce a message from other analyses — that is, the number of observations that go into constructing a QI for a given facility, controlling for the number of beds, may reflect differential censoring that has to be considered.

Having a cancer diagnosis was only weakly related both to the probability that the resident would be reported as having pain at a future assessment and to the probability that the patient would have been hospitalized or died. In addition, cancer is not uniformly distributed from state to state in the data base, nor from facility to facility in any given state.

Furthermore, there was no relationship detected between diabetes and the incidence of censoring or of incident pressure ulcers. Indeed, differences in the censoring rate associated with such casemix differences were relatively small, on average. A patient-level multiple logistic regression model applied to the data could explain only one percent of the variation in the rate of follow-up in the data base. Thus, the primary conclusion to be drawn is that there are some, but not all, outcomes that may be biased by differences in the censoring rate due to death. And, even when present, these relationships tend to be relatively weak. More specific details on these analyses can be found below.

Figure A.1.2
One-year resident loss distribution among 1,500+ nursing home facilities



Cancer is known to be a risk factor for daily pain among nursing facility residents (Bernabei, et al, 1998). At the same time, cancer is also a predictor of short-term mortality. Table A.1.2 summarizes the relationship between cancer and future pain among long-stay residents who were not rated as being in pain at the baseline assessment.

Table A.1.2
Relationship Between Cancer and Pain at Follow-up

Daily Pain at Followup	Has cancer	Does not have cancer
No	70.70%	77.40%
Yes	16.10%	12.90%
Died	7.50%	4.50%
Hospitalized	1.60%	1.20%
Missed or Late Assessment	4.10%	4.00%
TOTAL	100.00%	100.00%

As can be seen, a diagnosis of cancer is somewhat related to the "incidence" of daily pain as well as to mortality, but not to hospitalization as the reason for not having had an assessment.

Table A.1.3 presents similar information for the combination of diabetes and development of a pressure ulcer at follow-up. In this instance, while diabetes is related to the risk of future pressure ulcers, it is not related to either mortality or hospitalization.

Table A.1.3
Relationship Between Diabetes and Pressure Ulcers at Follow-up

Pressure Ulcer at Follow-up	Has diabetes	Does not have diabetes
No	85.2%	82.7%
Stages I-IV	4.3%	6.0%
Died	4.4%	4.5%
Hospitalized	1.4%	2.1%
Missed or Late Assessment	4.7%	4.7%
TOTAL	100.0%	100.0%

Since variation in the rate of censoring by state provides a proxy for the rate of censoring from facility to facility within state, we examined the relationship between state and the likelihood that the resident had a measure at follow-up. We found that among long-stay residents, 87.7 percent of Kansas residents had follow-up assessments. In New York, 88.2 percent of long-stay residents were assessed in time to meet the follow-up. As seen above, few hospitalizations occur that prevent timely follow-up assessment. Nonetheless, the prevalence of hospitalizations in a 90-day period does vary by state and may have an influence on the outcome measurement, thereby affecting the QI. In Mississippi, 14.1 percent of long-stay residents are hospitalized in a 90-day period, whereas in Maine only 5.7 percent are hospitalized and in New York, 9.4 percent of residents are hospitalized.

Multiple logistic regression analyses were conducted to identify the inter-relationship among the possible predictors of a resident not having a follow-up assessment, regardless of the reasons (Table A.1.4). The results are presented below. While the diagnostic variables are indeed predictive of not having a follow-up assessment, perhaps the most important fact to consider is that less than one percent of the variation in the outcome variable is accounted for by the clinical, casemix variables in the model. Numerous other resident-level factors were considered without much increase in explanatory power. Similar models were executed, all seeking to reduce the unexplained variation in hospitalization rates. All had similar results. In essence, without taking into consideration facility factors, such as resources invested, staffing, and ownership, there is no substantial increase in explanatory power. Even taking such factors into consideration, the explanatory power is not large, suggesting that there is considerable idiosyncratic heterogeneity in the rate of follow-up and discharge, even among long-stay residents. This is a mixed message and the lack of a strong correlation is a positive factor as we contemplate broad scale use of the QIs.

A.1.4 Summary of Evaluation of Censoring Differences

These analyses of variations in the censoring rate of residents across facilities reveal that different types of facilities, admitting different types of patients, have somewhat different QI "outcomes", partially as a function of which patients remain in the facility for follow-up assessments. Facilities wishing to demonstrate superior quality could arrange to hospitalize, or otherwise discharge, patients who have deteriorated. Fortunately, we find little evidence for the wholesale presence of this phenomenon. Earlier studies and detailed examinations of these data have revealed relatively little inter-facility transfer, either directly or by way of an acute hospitalization (Mor, et al, 1997; Hirth, et al, 1999).

The analyses of both data sets suggest that variation in the censoring rate is related to the type of facility. Censoring is higher among facilities concentrating in Medicare patients, as well as among facilities caring for dying patients. This variation in facility purpose is only partly accounted for by the mix of patients observed in the facilities studied, since our regression models did not explain a lot of the variation in discharge patterns. This suggests that there will be some "incorrect" QI rankings because facilities that have a practice of keeping patients when they deteriorate rather than hospitalizing them will appear to perform worse than other facilities that are otherwise similar. Future work on the development and interpretation of QIs will have to take these types of practice pattern differences into consideration.

Table A.1.4
Results of Logistic Regression Model Predicting Long-Stay Residents' Follow-up Status

Variable	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
Male	.2787986	.0292628	9.527	0.000	.2214447	.3361526
cxfever	-.6313881	.0751752	-8.399	0.000	-.7787287	-.4840474
Unstable	-.1760764	.057031	-3.087	0.002	-.2878552	-.0642976
Alzheim	.0970009	.0400799	2.420	0.016	.0184457	.1755562
Cancer	-.2617114	.0472507	-5.539	0.000	-.3543211	-.1691017
Congestive heart failure	-.2382575	.0315181	-7.559	0.000	-.3000319	-.1764831
Diabetes	-.0604347	.0341520	-1.770	0.077	-.1273714	.0065021
Pneumon	-.4063037	.1095772	-3.708	0.000	-.6210711	-.1915362
PVD	-.1500874	.0432828	-3.468	0.001	-.2349201	-.0652548
Cognition	-.0274914	.0079284	-3.467	0.001	-.0430308	-.0119519
Mobility	-.0935847	.0087328	-10.716	0.000	-.1107007	-.0764687
ME	.2792152	.1101775	2.534	0.011	.0632714	.4951591
MS	.3140099	.1070596	2.941	0.003	.105057	.5247220
NY	.0646076	.0999783	0.646	0.518	-.1313462	.2605614
SD	.6816195	.1178941	5.782	0.000	.4505514	.9126876
Intercept	1.907206	.1112468	17.144	0.000	1.689166	2.125245

A.2 Misclassification of Clinical Conditions – Evaluating the Presence of Ascertainment Bias

Ascertainment bias occurs because of variation in the ability of nursing facility staff to perceive the complex clinical dynamics inherent in the assessment of resident status in ways that are consistent with the MDS item instructions. This problem may be most severe in domains that are more difficult to assess clinically, e.g., delirium, mood distress, pain, recent weight loss, and dehydration. This means that staff in some facilities have both the necessary skills and assessment protocols and are more diligent in noting the presence of clinical problems for which intervention strategies can then be implemented. Such facilities will appear to have higher than normal proportions of residents with conditions that might trigger a QI. The key phrase is "appear to have," for the differentiating factor is not one of problematic care but rather assessment acumen.

Should those differences in coding skills be unrelated to the quality of care provided in a facility, they would only lead to random errors in the QI rates as our measure of quality. The conclusions that we could draw from QIs would thus be less precise, but still unbiased. The more likely assumption, however, is that facilities with excellent coding skills provide better care. If this were the case, we would systematically overestimate adverse event rates in those facilities resulting in ascertainment bias.

A.2.1 Analytic Approach to Evaluate Ascertainment Bias

A two-step analytic process was followed to determine whether concerns for ascertainment bias were justified. In step one, several candidate measures of quality performance derived from the first version of the MDS (1.0) were analyzed. As described below, some facilities with much higher rates of conditions such as pain and mood distress could not be explained using reasonable measures of casemix. We then used MDS 2.0 data to test whether the introduction of improved measures in the areas of pain and mood reduced the prevalence of outlier facilities in these measurement areas. The newer measures of pain and mood disturbance considerably reduced the degree of inter-facility variation.

Analysis One Relative to Ascertainment Bias. To test the proposition that differential ascertainment exists and influences the QI "scores" of a facility, a four step process was followed using Version 1.0 MDS data: 1) the identification of an appropriate resident cohort; 2) the selection of a sample of facilities; 3) the selection of candidate measures where differential ascertainment was most likely to be observed -- in this case in the areas of pain and mood distress; and 4) the specification of ordinary least square (OLS) regression models to determine whether facilities with high rates can be "explained away" by the unique mix of patients admitted into the facility.

We first sought to identify residents whose status at the time of the assessment was unaffected by care at the facility. Observed inter-facility differences in the average status of residents could be attributed to either differences in the measurement skill of staff or to differences in patient casemix at the time of the assessment. We sought residents where this ambiguity did not exist. New admissions meet these conditions. They generally have been in the facility for less than eight days at the time of their baseline assessment; their care plans are just beginning to be formed; and their status is largely unaffected by intervening care processes at the facility. As such, differences between new residents at different facilities are presumably attributable to either measurement differences or casemix differences and not to the pattern of care provided by the facility.

We considered several candidate measures to test the proposition that ascertainment affects the aggregated QI score constructed for a facility: pain, mood distress, weight loss, and delirium, to mention a few. In each case, we observed a reasonably long tail to the prevalence of the measure among facilities even though the overall average rate of the phenomenon across all facilities was at a much lower level. Table A.2.1 displays these data -- for mood and pain, as well as for delirium and weight loss.

Table A.2.1
Distribution of MDS Assessment Items that Might Fit the Ascertainment Bias Model

Areas of Assessment	Percent of Facilities With No Residents With the Problem	Percent of Residents Who Exhibit the Problem in the Average Facility	Percent of Facilities in Which 30+% of Residents Have the Problem	Percent of Facilities in Which 40+% of Residents Have the Problem	Percent of Facilities in Which 50+% of Residents Have the Problem
Residents with 2 or more mood problems	9.0%	16.1%	16.8%	8.4%	4.7%
Residents complain of pain	1.4%	25.5%	36.1%	16.4%	6.3%
Any delirium symptom	14.1%	11.7%	8.8%	3.0%	1.5%
Recent weight loss	9.4%	16.2%	17.6%	6.4%	0.7%

Our analyses concentrated on mood distress and pain, since previous analyses of these data in the published literature found considerable inter-facility variation, high prevalence of inadequate treatment and a tendency of staff to “normalize” these symptoms in the nursing facility population and, therefore, not identify their presence (Bernabei et. al., 1998). The unit of analysis is the facility. The dependent variable is the proportion of residents rated as having two or more mood symptoms and/or who were reported to be in daily pain. The primary independent variables reflect the proportion of residents in the facility admission cohort with indications of the clinical conditions that have been shown to be related to these two symptoms (e.g., measures of cognitive status, functional status, medical instability). Explicitly excluded were measures that are very close in concept to the measure being modeled (e.g., psychiatric diagnoses for the mood equation).

Table A.2.2 displays the OLS model for residents with two or more mood problems. Table A.2.3 displays the model for daily pain. For mood, 27 percent of the cross-facility variation in the average proportion of admitted residents with mood distress is explained by six casemix variables. Facilities with higher proportions of residents with multiple QIs of mood distress have a mix of residents at entry that is more likely to have medical instability and shortness of breath, poor vision, restricted decision-making ability and absence of contact with family or friends.

In the case of pain, 43 percent of the cross-facility variation in the average proportion of admitted residents rated as experiencing daily pain is explained by seven casemix variables, including poor health status as indicated by medical instability, constipation, falls and cancer, less restricted cognitive skills and lower levels of bladder incontinence.

Table A.2.2
Ordinary Least Square Model for Residents With Two or More Mood Problems (Standardized Regression Coefficients greater than .10 included)

Independent Covariates	Zero Order Correlation	Standardized Regression Coefficients [all significant at .001 or lower]
Medically unstable	.35	.18
Shortness of breath	.30	.17
Decision making skills – average	.16	.24
Visual appliances	.26	.25
No visual limitation	-.25	-.13
Absence of contact with family or friends	.10	.11
Explained variation (R2) = .27		

Table A.2.3
OLS Model for Residents With Pain
(Standardized Regression Coefficient greater than .10 included)

Independent Covariates	Zero Order Correlation	Standardized Regression Coefficients [all significant at .001 or lower]
Alzheimer's diagnosis	-.25	-.15
Other dementia diagnosis	-.41	-.22
Medically unstable	.30	.18
Constipation	.37	.22
Fell in past 30 days	.35	.21
Cancer diagnosis	.23	.12
Bladder incontinence	- .39	-.19
Explained variation (R2) = .43		

Table A.2.4 displays the raw and adjusted distribution of facilities with high rates of mood distress and pain. The adjusted results are based on the OLS models. In both instances, in facilities with the most extreme scores (representing nursing facilities in which 50 percent or more of the residents have the indicated problem), we see a dramatic shift in the quality problem rates.

The prevalence of facilities in which 50 percent or more of residents experienced two or more mood symptoms went from 4.7 percent, to only 0.1 percent after adjustment. Thus, for mood, 98 percent of the extreme outliers now have a more moderate score. Applying these rates to a state with 500 nursing facilities, we would go from 24 facilities with extreme scores to only one facility with an extreme score.

The results for pain follow this same pattern. For pain, the percent of facilities with 50 percent or more of the residents in daily pain dropped from 6.3 percent of the facilities to only 0.3 percent. Thus, for pain, 95 percent of the extreme outliers now appear to be less extreme. Applying these rates to a state with 500 nursing facilities, we would have gone from 32 facilities with extreme scores to only two facilities.

Table A.2.4
Raw and Adjusted Distribution of Facilities With High Rates of Mood Distress or Pain
(N = 1508 facilities)

Areas of Assessment	Raw or Adjusted Values	Percent of facilities in which 40+% of residents have the problem	Percent of facilities in which 50+% of residents have the problem
Residents with 2 or more mood problems	**Raw**	**8.4%**	**4.7%**
	Adjusted	**0.3%**	**0.1%**
Residents complain of pain	**Raw**	**16.4%**	**6.3%**
	Adjusted	**5.4%**	**0.3%**

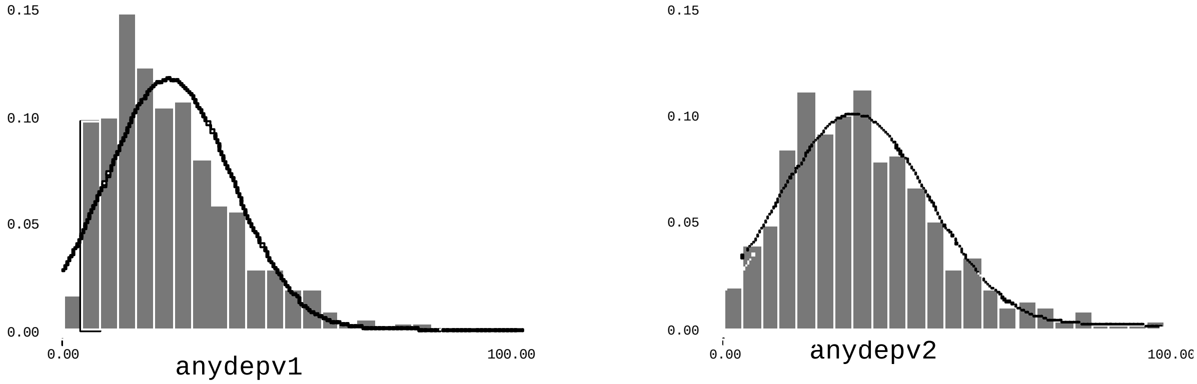
Obviously, differential ascertainment can still be present even at the “lower” end of the prevalence distribution and can never really be statistically adjusted away. By using the proportion of patients admitted to the facility with the problem (e.g., pain or distressed mood) or who are at risk of these conditions, it is possible to “adjust” for the observed prevalence of the problem in the resident population measured months later since the admission assessment provides a measure of how intensively the facility tends to look for the clinical problem. The results of this initial round of analyses should reduce significantly the disparity in quality rankings such facilities would receive.

Analyses of Ascertainment Bias using MDS 2.0 data. The above analyses were based on MDS Version 1.0 assessment data. Once MDS Version 2.0 data from CMS’s national MDS repository became available, additional analyses were conducted to determine whether ascertainment bias may have been reduced by the introduction of a superior measurement tool.

Figures A.2.1 and A.2.2 show the shift in distribution of the prevalence of both mood and pain QIs in long term care facilities in New York from 1996 (MDS V1.0) to 1998 (MDS V2.0). The mean prevalence for mood distress was 23.2 percent (median 20.5 percent) in 1996 whereas in 1998 it to rose to 30.2 percent (median 28.2 percent). Facility prevalence of pain showed a larger increase from 16.9 percent (median 15.4 percent) to 33.5 percent (median 32.1) in 1998. As in the previous analyses, only new admissions were included. Facilities were represented only if they had a minimum of 20 admission during the period. The additional analyses were encouraging in that differential ascertainment may not be as large a factor for potential bias as previously feared. While still a possible issue, once staff were given better items and directions for assessing the conditions, the ascertainment problem was reduced.

Figure A.2.1

Facility-level prevalence of any mood symptoms in New York State long term facilities with at least 20 admissions in 1996 (version 1, n=710) or 1998 (version 2, n=670).



Version 1

Mean 23.231

Median 20.463 Maximum 78.889

Std dev 13.637

Version 2

Minimum .559 Mean 30.230

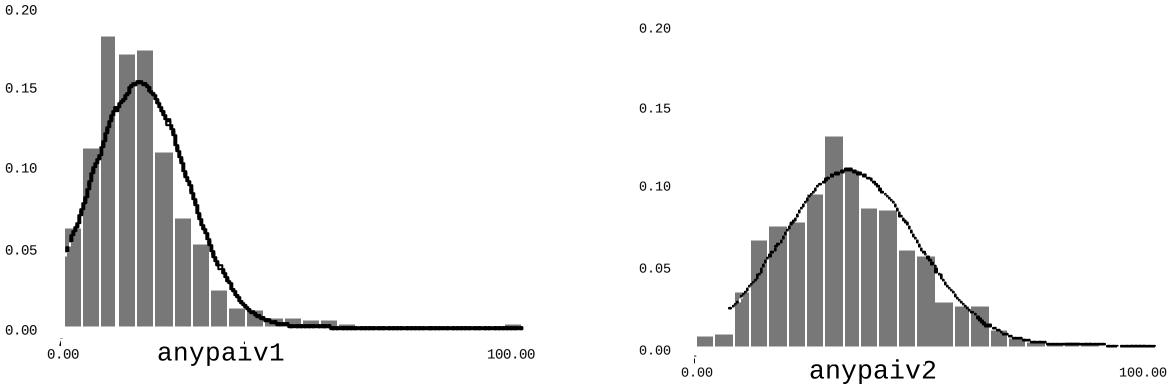
Minimum 338

Median 28.222 Maximum 92.857

Std dev 16.151

Figure A.2.2

Facility-level prevalence of any pain symptoms in New York State long term facilities with at least 20 admissions in 1996 (version 1, n=710) or 1998 (version 2, n=670).



Version 1

Mean	16.878	Minimum	.000
Median	15.385	Maximum	100.000
Std dev	10.482		

Version 2

Mean	33.466	Minimum	.000
Median	32.099	Maximum	84.158
Std dev	14.600		

A.3 Measurement Problems due to Small Numbers, and Inadequacy of Ranks to Determine Inter-facility Differences

A.3.1 Problems Due to Small Numbers

The stability of QI rates within facilities is partly dependent on the sample size, or number of residents in the facility during the quarter. Thus, it is difficult to characterize small facilities based on a single quarter of assessment data. A similar problem occurs for cross-sectional measures that are based on a risk-stratification model, in which residents at high- or low-risk of a problematic outcome are aggregated into a QI separately. Within a given 90-day period, few facilities will have sufficient cases to create usable QIs for both strata.

The following sections present analyses to examine the relationship between QI rates over successive time periods for cross-sectional and longitudinal QIs.

As shown in Table A.3.1, the consistency of cross-sectional QIs over successive time periods is higher than the consistency of change-based QIs, as they tend to reflect the prevalence of conditions which change little over consecutive 90-day periods. However, there is a considerable person-specific rather than facility-induced component to cross-sectional QIs for adjacent time periods — the majority of residents will be present at both measurement points and very few of these residents will have changed over the intervening time period. Thus, for these measures there is a tendency to have an artificially high inter-period correlation precisely because the same individuals are represented in successive QI measures and their condition has not changed. The natural rate of change to be expected of residents must be factored into the process. There is also a need to consider the state of the residents when admitted into the facility; if a facility admits a more impaired cohort, it will likely continue to appear to have a poorer outcome profile when in fact this was just a matter of adverse selection.

Table A.3.1.
Correlation Coefficients for QIs Assessed over Two Consecutive Quarters
(all facilities with 10 or more residents in denominator included)

	Cognition	Communication	Bowel	Locomotion	ADL	Bladder
Cross-sectional	.91	.91	.92	.93	.90	.90
Longitudinal	.39	.41	.42	.40	.42	.39

A.3.2 Analyses to Examine Rates of Change in Nursing Facility Residents

To understand how residents change over time, we have reviewed the cross-sectional and longitudinal performance profiles for three MDS-based summary scales: the MDS ADL Hierarchy scale; the Cognitive Performance Scale (CPS); and the MDS Communication scale. Each scale has a score ranging from 0 to 6, where zero is independent and 6 is dependent. All three have been reported in the literature (Morris, et al., 1994; Morris, et al., 1993; Phillips et al, 1997), they are all reliable, and they have been cross-referenced to external criterion.

The Activities of Daily Living (ADL) Hierarchy uses four measures of early, middle, and late loss functional performance to place residents into an ordered progression of increasing dependency. The CPS uses five cognitively-related items to place residents into an ordered hierarchy that has been shown to be related to scores on the Mini Mental Status Exam. The Communication scale brings together the two MDS items that tap expressive and receptive skill items (each of which is measured on a four-point scale, going from independence (a score of "0") to total dependence (a score of "3"). When summed together, these two communication items have been shown to result in a scale with high internal consistency (alpha reliability = .91).

For each of these scales, a decline in status is indicated when the latter summed score at the 90-day follow-up (e.g, Point A) is higher that the summed score at the earlier baseline measurement point (e.g., Point B). Residents who were totally dependent at baseline (e.g., Point B) and could not further decline were excluded from the aggregated change score calculation at the facility (i.e., their change measure value is set to missing). The proportion at any one time who are at the bottom end of the scale (where no further decline would be possible) are as follows: ADL Hierarchy, 17 percent; CPS, 15 percent; and Communication, 10 percent.

Over 90, 180, 270, and 360-day periods, decline and improvement rates are displayed in Table A.3.2. These data cover periods from three to 12 months, and the baseline period for each quarter is based on approximately 176,000 residents assessments, drawn from a seven-state data base. For the initial 90-day assessment period, the vast majority of residents are unchanged from the beginning to end of this period; 77 percent for ADL, 82 percent for CPS, and 84 percent for Communication. Even over a 12-month period, for residents who remain in the nursing facility for this entire period of time, most do not experience a change in status in these three areas: 58 percent experience no change for ADL, 66 percent for CPS, and 69 percent for Communication. Approximately six out of 10 residents who survived for 12 months were unchanged — their status at the end of the period was the same as their status at the beginning of the period.

For those who do change, in the initial 90-day period, 11 percent to 15 percent of all residents decline, while a somewhat smaller number improve – five percent to eight percent. Over time, the ratio of those who decline to those who improve becomes even more distinct. By 12 months, the proportion who have declined more than doubles – from 24 percent to 30 percent; while the proportion who have improved goes up at a much lower rate – averaging only about nine percent of residents by the one-year follow-up.
To examine how residents who declined or improved in the current time period faired in the earlier time period, we also looked backwards, starting with what happened to residents over the prior 90-day period. We found that for residents who declined over the most recent 90-day period, most had not declined in the preceding period. In fact, only 11 to 12 percent had so declined; while 10 to 15 percent had previously improved and about three-quarters (73 to 79 percent) had been stable over the earlier time period. Similarly, but with its own variation, for those who recently improved in status, a majority had been stable over the prior 90-day period: 57 percent for ADLs, 69 percent for CPS, and 67 percent for Communication. In addition, few of the residents who improved in the current quarter were following up on an improvement in status that first commenced in the prior quarter: eight

percent for ADL, six percent for CPS, and five percent for Communication. In fact, there is a significant number of residents whose current improvement is a rebound from a prior decline: 36 percent for ADL, 25 percent for CPS, and 29 percent for Communication.

Table A.3.2
Change Rates over One Year

Outcome Area	***Percent Who Declined***	***Percent Who Improved***	***Percent With No Change***
***ADL Hierarchy* – Change over 90-Day Periods**			
Change over initial 90 day period	**14.7%**	**8.2%**	**77.1%**
Change over 180 day period	**21.9**	**10.5**	**67.6**
Change over 270 day period	**26.4**	**11.8**	**61.8**
Change over 360 day period	**30.2**	**11.8**	**58.0**
***Cognitive Performance Scale* (CPS) – Change over 90-Day Periods**			
Change over initial 90 day period	**12.4%**	**5.3%**	**82.3%**
Change over 180 day period	**18.7**	**7.0**	**74.3**
Change over 270 day period	**22.7**	**8.1**	**69.2**
Change over 360 day period	**26.0**	**8.1**	**65.9**
***Communication Scale* – Change over 90-Day Periods**			
Change over initial 90 day period	**11.3%**	**4.9%**	**83.8%**
Change over 180 day period	**17.1**	**6.4**	**76.5**
Change over 270 day period	**20.7**	**7.6**	**71.7**
Change over 360 day period	**23.8**	**7.6**	**68.6**

Thus, for both decline and improvement, a clear majority of residents who have had such a current change in status come from among those who had been stable over the prior measurement period. For all residents, the most frequent condition over any one quarter, or even for any one year, is a lack of change.

For change-based measures we were concerned that no one time period would provide a sufficient sample on which to base the QI. For example, if one begins with a typical 80-bed facility, in which 65 of the beds are for long-stay residents and 15 for short-stay, post-acute residents, the expected number of cases that would be available for a QI would be determined as follows:

$ All post-acute residents would be lost to follow-up (n=15);

$ About 9 percent of the beds would be empty at baseline due to the absence of a full census at any one point in time – i.e., the typical facility in all states has less than full occupancy at any one point in time (n=6);

$ About 4 of the long-stay residents who were in the facility at baseline will have been discharged over the ensuing 90-day period due to death or relocation; and

$ About 15 percent of the remaining residents (n=8) will not be eligible for inclusion in the decline measure because they would have been totally dependent at baseline.

Thus, after these exclusions are considered, for the typical 80-bed facility, the aggregated change measure will be based on the experience of 47 residents. Using these rates of sample loss, Table A.3.3 displays a series of estimates of the expected samples that would be available to calculate a facility's QI score based on the number of residents who experienced a decline in Communication. [Note – the estimates for ADL and CPS would be about the same]. These estimates are displayed as a function of the size of the facility. We also include estimates of the number of residents in these samples who can be expected to decline based on the mean rate for the average facility, as well as the mean rates that were observed at given percentile points across the range for all facilities in the available five-state data base — at the 20th percentile point the facility rate equals four percent decline in Communication, at the 80th percentile point the rate equals 13 percent decline.

Table A.3.3
Estimates of Expected Samples

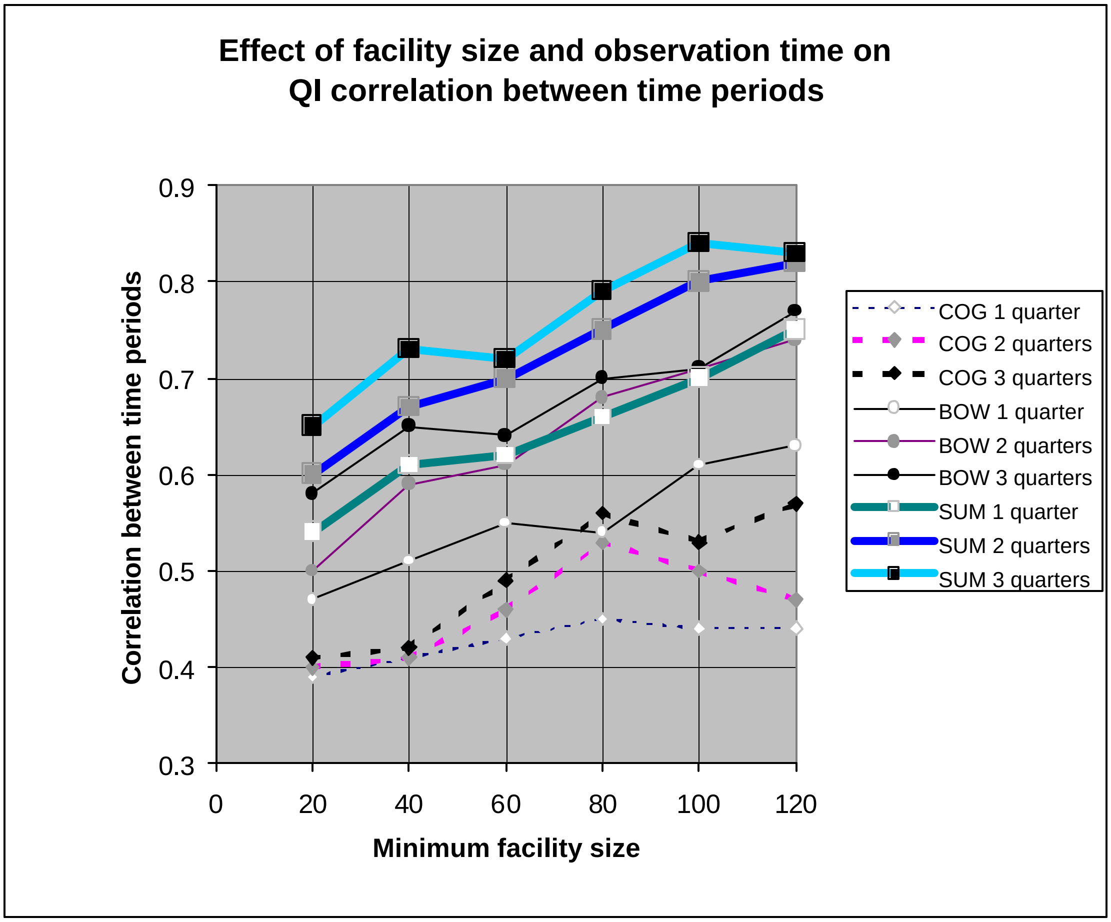
Size of Facility	Estimated Number of Residents on Whom the 90-Day Communication Decline Estimate Would be based	Expected Average Number of Residents Who Will Decline in Communication Over a 90-Day Period	At the 20th Percentile -- Expected Number of Residents Who Will Decline in Communication Over a 90-Day Period	At the 80th Percentile -- Expected Number of Residents Who Will Decline in Communication Over a 90-Day Period
20 Beds	12	1	<1	1
30 Beds	16	1	<1	3
50 Beds	28	2	1	4
80 Beds	45	4	2	6
100 Beds	56	5	2	7
150 Beds	83	7	4	11
200 Beds	117	9	5	14

A.3.3 Analyses to Show the Effect of Facility Size and Observation Time on the Correlation of QI Rates Across Time Periods.

One of the issues to be considered in assessing the validity of a QI is the stability of a facility's estimated QI value over consecutive time periods. As indicated earlier, the relationship between a facility's past and future will be imperfect — a certain range of variation can be expected. Measurement error, real shifts in facility performance, and natural deviations around some true score performance standard are among the factors that can contribute to such imperfect longitudinal relationships. The difference in the values of a QI over two consecutive measurement points needs to be addressed, and our goal is to identify measurement procedures that will reduce the random variation and thereby create more stable QI estimates. Through this process we seek to achieve two inter-related objectives. For the past, we seek to arrive at an estimate that best reflects actual facility performance in the recent past, reducing the likelihood that a facility's performance was inappropriately assigned too high or low a score based on errors of measurement or the unique experience of one or two residents. For the future, we seek an estimate such that once it is made (based on the recent past), it will prove to have reasonable prognostic value into the future. From a long term care survey perspective, the facility that did poorly in the past should have a reasonably high likelihood of continuing to have problems into the future. Such a facility may therefore continue to warrant special oversight. While from a resident-family perspective, if the decision to enter a facility was based on prior superior facility performance, one would hope that there would be a reasonable likelihood that such performance would continue into the future.

Cognitive and bowel QIs were calculated based on adjacent single quarters, adjacent two quarters and adjacent three quarters. A summary QI based on six functional QIs (cognition, communication, locomotion, bladder and bowel incontinence and ADL decline) was calculated for the same time frames. Figure A.3.1 shows that larger facilities tend to show greater consistency in rates from period to period. More striking was the consistency across time when assessment data from more than one quarter were used to calculate the QI rate within the facility.

Figure A.3.1
Correlation coefficients between the values of a QI calculated at two consecutive periods of time



A.3.4 Inadequacy of Rank Differentiation

QI rates are commonly used to rank facilities within regions. In reviewing the distribution of QI rates within states for existing QIs, we observed distributions that were often skewed or rates that were bunched closely together. Often the differences between one ranking and another had no practical meaning. Ranks do not represent unequal differences in rates. Depending on the distribution of the QI rates, facilities that ranked within five places may differ in rate by 0.05 percent whereas at a different point in the distribution a difference of five ranks may correspond to a difference in rate of greater than 25 percent. In many cases, ranks will make it appear that homes are different from one another in quality when, in fact, they are clinically and statistically indistinguishable.

A.3.5 Summary Relating to Small Numbers and Inadequacy of Ranks

QIs must overcome the technical issues of small numbers both in the numerator (as illustrated by the small numbers who change within a given quarterly period) and in the denominator (the number of residents who are at risk of developing the problem). A possible solution to these issues include increasing the number of quarters of assessment data used to calculate the QI rate. Additionally, when developing new QIs, it is advisable to refrain from addressing problems that are rare.

Given the varying distributions of QIs, it is not clear in advance which constitute meaningful thresholds that reflect poorer or better care. Rankings are particularly susceptible to misinterpretation as QIs of quality differences. It will be important to set clinically meaningful as well as statistically meaningful thresholds to guide decisions regarding better or worse performing nursing facilities.

A.4 Multi-dimensional Nature of Nursing Facility Quality

As noted in the earlier report, existing QIs characterizing nursing facilities cover a wide variety of domains from specific clinical conditions such as pressure ulcers to global phenomenon such as decline in ADL. The following section addresses the issue of whether there is a pattern of inter-relationships among QIs. Table A.4.1 provides correlation coefficients for the 26 provisionally accepted QI variables. They are organized into five domains: Functional Decline, Mood/Behavior, Pressure Ulcers, Treatments, and Conditions. Only correlations of at least .20 were included in this table. This section of the report describes the process by which these 26 QIs were grouped.

Initially, the 26 QIs were entered into a factor analysis which employed a varimax rotation. This analysis established five factors.

> *Factor 1* was composed of four longitudinal (change) QIs: bladder (LTCQ), bowel (LTCQ), activities of daily living (CHSRA), and locomotion (LTCQ).
> *Factors 2 and 3* include both prevalence (cross-sectional) and longitudinal (change) QIs. *Factor 2* was composed of the following six QIs: mood (CHSRA), mood/no antidepressants (CHSRA), mood (LTCQ), high-risk behavior (CHSRA), behavior (LTCQ), and relationships (LTCQ). *Factor 3* was composed of the following three

QIs: pressure ulcers (LTCQ), high-risk pressure ulcers (CHSRA), and low-risk pressure ulcers (CHSRA).
Factor 4 was composed of two QIs: cognition (LTCQ) and communication (LTCQ), both longitudinal change QIs.
Factor 5 includes low-risk bladder/bowel, feeding tubes, falls and pain.

Upon review of the correlation matrix in Table A.4.1, we observed that the correlation between cognition and communication was relatively strong (.64). We reasoned that because these QIs tended to have moderately strong correlations with QI measures subsumed under Factor 1, one or both of these QIs might join Factor 1 if not influenced by the other QI. We confirmed this interpretation by repeating the analyses with and without the measures of cognitio n and communication respectively. As suspected, each of these two QIs, apart from the influence of the other QI, were grouped into Factor 1.

The remaining QIs were not included in any factors. However, we loosely classified them as either a treatment or a clinical condition. Treatments include catheter (CHSRA), restraints (CHSRA), hypnotic (CHSRA), low risk antipsychotic (CHSRA), and high-risk anti-psychotic (CHSRA). Conditions include weight loss (LTCQ), falls (LTCQ), pain (LTCQ), and low-risk bladder/bowel (CHSRA). The only exception was low-risk behavior that could not be considered a treatment or condition. Because of its moderately strong correlation with high-risk behavior (CHSRA) (r= .55) and behavior (LTCQ) (r= .31), it was included in the Mood/Behavior domain.

Summary of Multi-dimensionality analyses

The results of the factor analyses do not support a uni-dimensional notion of facility quality of care. We identified five latent factors which we named functional decline, mood/behavior, pressure ulcers, treatments and conditions. It is possible that the factors may represent yet other constructs. For example, the functional factor may relate either to function as a construct or to decline among residents.

Table A.4.1
Correlation Coefficients for 26 QIs

		FUNCTIONAL DECLINE						MOOD/BEHAVIOR Decline (D) and Prevalence (P) Measures						
		Bladder (LTQ)	Bowel (LTCQ)	ADL (CHSRA)	Locomotion (LTCQ)	Cognitive(LTCQ)	Communication (LTCQ)	Mood (CHSR A)	Mood/No Antidepressants (CHSRA)	Mood (LTCQ)	Behavior High-Risk (CHSRA)	Behavior (LTCQ)	Relationship (LTCQ)	Behavior Low-Risk* (CHSRA)
Functional Decline	**Bladder (LTCQ)**		**.58**	**.43**	**.44**	**.35**	**.39**			**.24**		**.26**		
	Bowel (LTCQ)	**.58**		**.44**	**.43**	**.32**	**.35**			**.23**		**.28**		
	ADL (CHSRA)	**.43**	**.44**		**.68**	**.38**	**.31**			**.32**		**.27**		
	Locomotion (LTCQ)	**.44**	**.43**	**.68**		**.37**	**.32**			**.32**		**.26**		
	Cognitive (LTCQ)	**.35**	**.32**	**.38**	**.37**		**.64**			**.32**		**.26**		
	Communication (LTCQ)	**.39**	**.35**	**.31**	**.32**	**.64**				**.30**		**.23**		
Mood/ Behavior	**Mood (CHSRA)**								**.93**	**.37**	**.39**		**.40**	
	Mood/No Antidepressants (CHSRA)							**.93**		**.32**	**.33**		**.36**	
	Mood (LTCQ)	**.24**	**.23**	**.32**	**.32**	**.32**	**.30**	**.37**	**.32**		**.29**	**.41**	**.30**	
	Behavior High-Risk (CHSRA)							**.39**	**.33**	**.29**		**.44**	**.30**	**.55**
	Behavior (LTCQ)	**.26**	**.28**	**.27**	**.26**	**.26**	**.23**			**.41**	**.44**			**.31**
	Relationship (LTCQ)							**.40**	**.36**	**.30**	**.30**			
	Behavior Low-Risk* (CHSRA)										**.55**	**.31**		
Pressure Ulcers	**Pressure Ulcer (LTCQ)**	**.26**	**.24**											
	Pressure Ulcer High-Risk (CHSRA)													

Table A.4.1
Correlation Coefficients for 26 QIs

		FUNCTIONAL DECLINE						MOOD/BEHAVIOR Decline (D) and Prevalence (P) Measures						
		Bladder (LTQ)	Bowel (LTCQ)	ADL (CHSRA)	Locomotion (LTCQ)	Cognitive(LTCQ)	Communi-cation (LTCQ)	Mood (CHSR A)	Mood/No Anti-depressants (CHSRA)	Mood (LTCQ)	Behavior High-Risk (CHSRA)	Behavior (LTCQ)	Relationship (LTCQ)	Behavior Low-Risk* (CHSRA)
	Pressure Ulcer Low-Risk (CHSRA)													
TREATMENTS Prevalence Measures	Catheter (CHSRA)													
	Restraints (CHSRA)													
	Hypnotic (CHSRA)									.22				
	Antipsychotic Low-Risk (CHSRA)													
	Antipsychotic High-Risk** (CHSRA)									.26	.20		.23	
	Feeding Tubes (Ramsey)			.20				.21		.32			.23	
CONDITIONS Decline and Prevalence	Weight Loss (LTCQ)									.20				
	Falls (LTCQ)				.24	.20				.23				
	Pain (LTCQ)			.24	.25	.22	.21	.21		.39		.22		
	Bladder/Bowel Low-Risk (CHSRA)	.26	.27											

* (n = 786)
** (n = 343)
Note: Correlation coefficients smaller than 0.20 were omitted.

A.5 Summary Interpretation of Supplemental Empirical Analyses

We conducted the above sets of inter-related empirical analyses in order to help us understand the degree of confidence that users of the recommended QIs could have in the resulting classification or rankings of the facilities on various measures of qua lity. The results of the "ascertainment bias" and the "selection bias" analyses clearly reveal that casemix differences help to explain a large amount of the variation in selected QI scores that facilities receive, but not all such scores. Regarding differential ascertainment, the project team found that facilities with a high prevalence of clinical outcomes, which are prone to miscoding, like pain, also differed from other facilities in the kinds of residents they admitted. Once the more precise definitions of MDS 2.0 have been introduced, a good deal of this problem may have been solved. In the case of differential censoring, we found that residents with clinical characteristics that predisposed them to daily pain had a slightly higher likelihood of death or being hospitalized before a follow-up MDS measurement and that this phenomenon was more prevalent among certain classes of facilities. In both cases, however, differences in QI scores remained after considerable casemix adjustment. Nonetheless, censoring is a phenomenon that should be monitored, but it is not sufficiently powerful or prevalent to undermine the utility of the measures. Indeed, our analyses suggest that the inclusion of a measure characterizing the mix of patients admitted into the facility as a modifier or adjuster of a facility's QI, will reduce the impact of selection, censoring and ascertainment bias.

Furthermore, while sample size concerns will always be present in making projections about the quality of a home, it is possible to incorporate more than one quarter of data into the construction of a QI, thereby increasing the stability of the QI and its ability to detect real differences in the rates of problems between facilities. Finally, it is clear that quality as conceptualized and empirically measured using existing QIs is fundamentally multi-dimensional, meaning that there will always be multiple measures of facility quality and facilities, surveyors, consumers and purchasers may have different preferences for which of these should be more or less influential in deciding how to act. Technical Appendix B describes a modest proposal under consideration for handling the technical and methodologic issues raised here.

Technical Appendix B: A Modest Proposal for Benchmarking QIs to Identify "Good" and "Bad" Facilities

This technical appendix summarizes many of the issues raised in Chapter 7 of this report and offers a modeling strategy for analyzing QIs that can potentially overcome those limitations. We present this approach, which to our knowledge has not been employed in prior research on QI development, using a universally accepted long-term care outcome of interest – the development of pressure ulcers. Although it clearly does not overcome all the identified conceptual and technical problems at this stage, we believe that this modeling strategy can be applied to the construction of QIs for other outcomes such as the array of QIs recommended in earlier chapters of this report.

Background

Depending on the targeted audience, QIs have to fulfill different requirements. The current report recommends the adoption of a series of existing QIs derived from resident level MDS assessment data, but feels most comfortable using these measures for quality improvement efforts a facility undertakes. Use of these QIs for the purpose of identifying facilities that merit additional scrutiny as part of the regulatory inspection process is also considered. However, the authors of this report believe that none of the existing QIs are sufficiently robust to meet the stricter requirements for such a regulatory rather than a managerial application. Modifications in either the construction of the QI at the resident level or in the adjustment of QI rates for casemix at the facility level appear necessary, before those QIs can be used by purchasers, consumers or regulators.

Caution about the adoption of existing QIs for the purpose of ranking facilities is based upon a variety of technical and conceptual concerns reviewed in this report. These can be broadly grouped into concerns about misclassification, i.e., that a facility will be classified as "good" when it is not, or classified as "bad" when it is not, and about fairness, i.e., that the strategy used to group facilities adequately accounts for casemix differences. In addition to misclassification and fairness, there are some basic issues about the meaning of quality in long-term care and what QIs can tell us about our shared understanding of quality. In the paragraphs below, we reiterate some of these conceptual issues and how we interpret them since the paradigm for the creation of a new type of QI we are proposing is based upon our understanding of these conceptual issues.

The Multidimensional nature of Nursing Facility Quality. Based upon analyses presented in Chapter 7 and Technical Appendix A, quality appears to **not** be a uni-dimensional concept. Some facilities perform poorly on some QIs of quality but perform well on others. Other researchers have obtained similar results in their analyses of QI data drawn from different states and using somewhat different criteria. Furthermore, structural variables commonly associated with nursing facilities quality, such as staffing levels, are commonly related to some but not all QIs. These two findings complicate any attempt to develop a single measure that can be used to characterize a provider as "good" or "bad".

As shown in Technical Appendix A, reducing the complexity or the number of different clinical domains of quality is not going to be helped by data reduction. Rather, some conceptual, or value based, approach is going to be necessary to ascribe greater or lesser importance to some measures of quality. If we focus on those aspects of quality about which there is little ambiguity as to their meaning and value, we could eliminate some of the inherent complexity in characterizing homes as being of “poor” or “good” quality. While still complicated, it is not so problematic to decide what is “good” and what is “bad” if we restrict ourselves to a relatively small set of measures about which there is substantial agreement as to their generalized salience. On the other hand, this reductionistic approach may not give sufficient credence or importance to some domains of quality that some people deem to be very important.

Unfortunately, when most of us naively conceptualize quality we tend to think of it as either “all or nothing” or, at best, as the level of overall quality. It is difficult to conceptualize quality as having as many dimensions as there are QIs (or close to it). If it is hard for specialists to conceive of homes as being “good” and “bad” at the same time, imagine how unusual this idea is to purchasers and consumers. It is hard to imagine that consumers would be able to discriminate preferences across so many domains simultaneously. Nonetheless, contrary to our expectations and desires, it does not appear to be possible to identify “good” and “bad” nursing facilities in a global sense based upon the QIs now in use. Whether this ambiguity about what constitutes quality is confusing to the surveyors and regulators is not known. However, all facilities engaged in continuous quality improvement processes obviously area aware that they meet their established standards in some areas and not in others. Indeed, the point of CQI is to identify those clinical areas that could benefit from improvement and those which have met the established performance criterion. Thus, good care providers behave as if quality is multi-dimensional, but the general public and consumers conceive of quality as uni-dimensional.

The implicit “meaning” of a QI. Related to the assumption of uni-dimensionality, QIs are inherently assumed to capture the present AND future quality of a facility. That is, regardless of the application to which a QI is put, we implicitly assume that it captures performance for the period covered by the measurement as well as the current situation in the facility. This means that QIs should be stable; that last quarter’s measurement, all things being equal, will predict next quarter’s measurement precisely because the QI captures something about the performance of the facility and not just a transitory measurement state. There are many reasons for wanting a QI to be stable. Providers don’t want to be in the bottom rank one month after having been in the top the month before. Consumers and purchasers selecting a facility based upon past performance imagine that the facility will continue to perform as it had to achieve its advertised QI score or rank. Obviously, there may be some circumstances in which radical changes in the organization can translate into rapid deterioration in quality performance, but we would imagine that these would be the exceptions and might even be identifiable based upon certain structural factors (e.g., change in senior leadership or wholesale staff flight). Thus, a QI should have a predictive aspects associated with it and not just a retrospective recording.

Indeed, the notion of continuity of performance is key to the utility of QIs no matter their application. In many respects a QI is not worthwhile if once cannot count on it making a statement about the future as well as the past. Even use of QIs for continuous quality improvement purposes requires that the measures are sufficiently stable that if the home does nothing to rectify the conditions that led to poor performance the QI would continue to reveal sub-standard performance. Thus, for both technical and conceptual reasons, QIs must tell us about the future and not just the past.

Different Measures and Standards for Different Audiences. We've been conceiving of our QIs as having four possible audiences: facilities' quality improvement interests, regulators' needs to identify poor performers, purchasers' interest in buying only from "good" homes and consumers' interest in avoiding "bad" homes and, if possible depending upon their location, finding good ones. The first option is always easiest since we all agree that no home is so good that it couldn't improve and shouldn't strive to improve. The second option is far more complex due to the regulatory authority surveyors have with its potential for financial fines. If QIs incorrectly single out a home to regulator as having a serious problem in an area then surveyors are likely to look particularly thoroughly for that problem and therefore will be more likely to find it than they might have otherwise. Nonetheless, at least surveyors have an opportunity to observe the facility and to be convinced that the QI report signaling the home as "poor" may have been incorrect. In contrast, purchasers and consumers never have the opportunity to disprove the validity of the QI since they make a purchasing decision based solely on the QI data with no recourse to other information.

As noted in Chapter 7, we separately rated the adequacy of each recommended QI for each audience. However, we made no effort to consider an alternate QI for each audience nor even an alternate format for summarizing, or collapsing, the same information about the quality of the home. For the most part, all existing QIs in use in the field end up in a relatively simple continuous ranking of facilities on the basis of a given QI. In some instances a threshold is established based upon the percentile rank (e.g., the top 10th percentile or the bottom 20th percentile). Depending upon the actual underlying distribution of the QI in question (e.g., pressure ulcer incidence or restraint use rate prevalence) it may be the case that a facility at the 20th percentile and one at the 80th percentile may have very similar scores. Establishing thresholds from rankings based upon QIs with such an underlying distribution might be quite volatile and certainly won't communicate the "actual" performance difference.

As noted, most QI systems rank homes on each QI based upon the “raw” value of the QI measure. This basically transforms into an ordinal measure the actual measure that the average resident (or a designated percentage of all residents in a home) experienced. For example, 4.1 percent of residents “at-risk” of acquiring a pressure ulcer may have been observed to have a PU 60 to 120 days after baseline in facility A. Let us say that facility B was observed to have a rate of 4.7 percent of incident pressure ulcers. In real terms there is no difference between these two numbers but they might be ranked 200th and 300th in a state with 500 facilities. Since underlying almost all distributions of QIs we’ve observed is either a skewed distribution, or one that is bunched in the center, transforming the raw scores into ranks creates more separation among the homes than is “real”. For QIs that look at change these distributions tend to be “bunched” between the top and the bottom third of the facilities. On the other hand, for measures like pressure ulcer incidence or restraint use, QI scores are skewed, with many facilities having close to zero events. In either case, an actual QI score will not differ dramatically among those at just into the top third and those just under the bottom third of homes. Using ranks to differentiate will make it appear that the homes are quite different from one another when, in fact, they are statistically indistinguishable.

An alternative to using “ranks” as the basis for classifying homes is to determine the “cut-points” that distinguish between “good” and not so good homes on a given measure. Setting these “cut-points” requires either a distributional or an a priori standard, or benchmark. Determining whether a facility is statistically significantly above or below the established cut-point can be done using statistical models based upon analyzing the actual “raw” variables.

It is certainly possible, using the same underlying QI score distribution, to have different cut-points, different foci (best or worse) and even different levels of confidence about the designation that a home is “good” or “bad”. The same scientifically valid QI measure, whether risk adjusted or not, whether averaged over multiple quarters or observations or not, could form the basis for alternate summary expressions of how to classify a home on the basis of a given domain. One of the purposes of this chapter is to offer such an alternative approach and some evidence for its statistical and conceptual viability.

Establishing Cut-Points Based upon Benchmarks. As will be described below, efforts to distinguish homes with high and low quality rankings suggest that this effort was not possible with any consistency. However, reconsideration of the task in light of the various audiences who might use QIs offer an alternative regarding classifying facilities as “good” or “bad”. For the home’s own quality improvement efforts, just knowing its’ score as well as the minimum and maximum possible score may be sufficient to motivate it to improve. For the regulator, a QI should identify facilities that are likely to have quality problems, but, since there is still an opportunity to see for oneself by surveying the facility, some degree of error in the classification may be acceptable. On the other hand, since purchasers and consumers have to make a irrevocable decision on the basis of the available information, the level of confidence should be higher for this application. The basis for establishing cut-points for each audience would be greatly improved were there established standards based upon large data bases, particularly if these were reinforced by clinically based “standards”. Proposed mechanisms for each audience is summarized below.

Facility Quality Improvement efforts are based upon the premise that all homes would like to improve their performance in domains of clinical care. The key to communicating QI information to facilities in such a manner as to stimulate them to improve and not to become defensive is to provide them with information about their performance relative to their peers. For this, the home should be provided with its actually observed rate of the event among residents with low, average and high risk. As basis for comparisons, it is useful for the administrator to know the rate of the home's competitors at the 50th percentile as well as at the 75th and the 90th percentile. By and large, if the QI is thought to be important by a home, most homes will want to improve their performance unless they exceed the 75th or 90th percentile.

Regulatory Reporting to Guide Inspections is one of the key new functions that QIs are supposed to address. To fulfil this function, one can establish a cut point to identify badly-performing facilities based on statistical criteria. For example, a distribution based standard might identify homes with a QI rate that exceeds the peer group average by at least one standard deviation. However, by giving surveyors facility-level QI reports or lists of individual residents that trigger the QI condition , their prior impression of the home will be altered and they may seek evidence to support the initial data without questioning whether the data might be correct.

Purchaser and Consumer Performance Benchmarks are meant to form the basis of decisions which are more irrevocable than either the quality improvement or the regulatory ones. Furthermore, since these measures need to be understood by a lay audience in order to be credible, the face validity of the standard should be as obvious as possible. This means that, to the extent possible, the standard should be based upon the literature, clinical validity and the distribution of the event of interest observed in a long-term care populations. For example, the quarterly incidence of pressure ulcers in a population of nursing facility residents with low risk for developing one should be close to zero. Therefore, it would be reasonable to establish a cut-point at one percent and identify those homes that have an incidence rate in excess of 1 percent estimated with a 90 percent confidence interval. Depending upon the adequacy of the model, those homes exceeding that level would likely be poor performing facilities with respect to pressure ulcer prevention. Similarly, the literature suggests that competent homes are able to keep the incidence of pressure ulcers under 5 percent to 10 percent even among residents at the highest risk. It would be possible to identify those homes that have pressure ulcer rates in excess of such a criterion, estimated using the 80 or 90 percent confidence intervals. The number of facilities that would be identified as having a high likelihood of performing worse than these standards at any given time would depend upon facility size, observed incidence, the number of resident quarters used to create the facility estimate and the distribution of residents' risk factors across facilities.

The inverse of the above benchmark construction might also be possible. That is, homes in which the average risk resident experiences less than a 1 percent pressure ulcer rate can be calculated with 90 percent confidence as a "good" home at pressure ulcer prevention. The resulting listing of facilities that might meet that criterion could be selected by consumers as meeting a relatively high standard of performance.

To summarize, current QIs tend to rely upon ranking systems to differentiate homes, in spite of the fact that rank based differentiation is known to find differences where none really exist.

A "Modest" Proposal for Overcoming Conceptual and Technical Limitations of Current QIs

In light of the various conceptual and technical issues that limit current nursing facility QIs, we propose an approach that begins to overcome these limitations. This section of the chapter describes the new methodology and the manner in which it overcomes existing technical limitations. This is followed by the presentation of preliminary results of applying this approach to one QI concept to one state.

As noted above, both conceptually and empirically, nursing facility quality is a multi-dimensional phenomenon. That is, we are not likely to find a single measure that will reasonably suffice to rank all homes on the basis of the quality of care they provide. This means that the proposed approach to creating QI models we are proposing must be done separately for each QI domain considered to be sufficiently important to merit a separate measure.
As enumerated in Chapter 7 and Technical Appendix A, the threats to validity that a QI must overcome include:

Being based on too few cases, both in the numerator or denominator **(sample size)**;
Differences in assessment effort, or problem ascertainment **(ascertainment bias)**;
Differences in the mix of residents a facility serves **(casemix)**;
The inability to make statements about the statistical confidence with which a facility is classified as being "better" or "worse" than another **(statistical precision)**;
Differences in the mix of residents that facilities admit, or which are referred to them **(selection bias)**; and
The inability of ranks *per se* to adequately characterize differences between facilities **(rank differentiation)**.

The proposed model addresses these limitations of existing issues as follows:

$ **Sample size.** Increasing the number of quarters of assessment data used to calculate the QI rate, thereby increasing both the denominator and the numerator.

$ **Ascertainment Bias.** Using a measure of facility level indicative of the extent to which a clinical problem is identified amongst all residents admitted to the facility.

$ **Casemix.** Using a parsimonious set of risk adjusters.

$ **Statistical Precision.** To both establish externally meaningful thresholds, or performance levels and estimate the statistical probability that a facility exceeds that threshold in the population of residents served over the period of observation.

$ **Selection Bias.** Create a measure characterizing the proportion of all resident admitted with the clinical problem, regardless of whether those residents remain in the facility.

$ **Rank Differentiation.** To create statistically meaningful comparisons between facilities that are significantly "better" than facilities in the bottom 25 percent or 10 percent of facilities in the distribution or "worse" than facilities in the top of the distribution.

As is apparent, these limitations affect the ability of a QI to fairly and statistically reliably classify a facility as being a good versus a poor performer. This has to do with the stability of a measure being based upon a reasonable number of residents which therefore means that the observed rate is unlikely to be spurious, or accidental. It also is related to the heterogeneous character of nursing facilities and the population of residents they serve. QIs should not be created in a way that penalizes providers for offering care to more severely ill residents. The issue of risk-adjustment is especially problematic in the long-term care industry. Many nursing facilities care for a wide range of acuity levels as they attempt to balance the provision of post-acute care with traditional long-term custodial care. Because of the unique position of nursing facilities along the continuum of care, risk-adjustment must go beyond the resident level to include some measure of the profile of residents admitted.

For example, in the case of pressure ulcers, facilities care for residents with varying levels of intrinsic risk (i.e., functional problems, diabetes, incontinence, etc.). Yet nursing facilities also inherit the results of the good and poor care practices of hospitals. The poor practices of hospitals are readily apparent in the profiles of nursing facility residents at admission. Based upon all MDS admission assessments in New York in 1999, the average rate of pressure ulcers recorded upon admission to the nursing facility was 18 percent.

Clearly this rate was not the same across all facilities in the state. Some nursing facilities specialize in admitting these more clinically complex discharges from hospitals, while others operate to discourage the admission of such residents. Either way, facilities tend to select residents, or to have them referred, from hospitals that closely match their resources, skills and mission. Since residents with a history of a pressure ulcer are significantly more likely to acquire one in the future (for physiological and even measurement reasons due to difficulties in reverse coding), facilities admitting residents with pre-existing pressure ulcers run the risk of looking worse on a pressure ulcer QI simply because they admit a higher acuity population at admission. This **selection bias** phenomenon can undermine the actual and perceived fairness of the QI comparisons. Facilities will have a disincentive to provide care for the most vulnerable if the QIs adopted by the government or accreditation agencies fail to properly adjust for this selection phenomenon.

Selection is closely related to **ascertainment bias** — the non-random way in which some facilities do a better job of identifying and measuring quality problems. Nursing facilities admitting a high percentage of residents with pressure ulcers are more likely to identify pressure ulcers at admission and at subsequent quarterly assessments and hence are likely to appear *worse* simply because they record more problems.

Preliminary Analysis: Justification for a New Approach

How adequately risk-adjusted are the existing pressure ulcer QIs?
One way to examine how well a QI accounts for the acuity of residents is to compare the QI against characteristics of nursing facilities associated with treating high acuity residents. Levels of skilled staffing should be inversely related to the prevalence and incidence of pressure ulcers if the QI properly adjusts for resident acuity.
In Table B.1, the relationship between staffing and the two most widely used pressure ulcer QIs (based on quarterly MDS assessment data) are presented. The relationships are either non-significant or modest in the opposite direction — more skilled staffing is associated with

worse performance on the pressure ulcer QI. In the case of the CHSRA prevalence measure, the "high risk" QI attenuates this negative relationship, but only slightly. The LTCQ incidence measure also fails to perform in a manner consistent with a properly risk-adjusted QI.

Table B.1
Partial Correlations Between CHSRA Pressure Ulcer QI and Staffing Characteristics (controlling for high Medicare volume or hospital-based status)

	Number of Facilities	Registered Nurse FTE per 100 residents		Total RN + LPN FTE per 100 residents	
		Correlation	P value	Correlation	P value
1999 CHSRA Prevalence Measure					
Low Risk[a]	541	.28	.000	.10	.023
High Risk[b]	542	.17	.000	.06	.182
1999 LTCQ Incidence Measure					
Unadjusted[c]	556	.08	.077	.02	.685
Adjusted[d]	556	.07	.081	.01	.984

Notes:
[a] Low risk = All residents excluding high risk at most recent quarterly assessment
[b] High Risk = Impaired transfer or bed mobility, coma, malnutrition or end stage disease at most recent quarterly assessment.
[c] Unadjusted = Total # of residents with a higher pressure score than on previous assessment dived by the total number of residents with two valid quarterly assessments
[d] Adjusted = Facility mean PU incidence times the grand mean PU incidence, divided by the average predicted probability derived from a logistic model with transfer, unstable, history of PU, open lesions, wound care, bed rails used, bowel incontinence, bladder incontinence, bed mobility problems and locomotion problems as covariates.

The counter-intuitive relationship between skilled staffing levels and performance on the CHSRA and LTCQ QIs also can be illustrated by examining the "worst" and "best" extremes of the QIs and comparing those to average staffing levels. In Table B.2, the “worst” performing 10 percent of facilities are shown to have a significantly higher number of skilled staff per 100 residents.

Table B.2
Relationship Between Staffing Levels and Worst/Best Performing Facilities Based on CHSRA and LTCQ QIs.

	Staff FTE Per 100 Residents					
	Worst 10% on QI		**Middle 80% on QI**		**Best 10% on QI**	
	RN	RN+LPN	**RN**	**RN+LPN**	**RN**	**RN+LPN**
	Mean (SD)	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**
1999 CHSRA Prevalence Measure						
Low Risk[a]	10.0 (8.6)	23.9 (13.0)	6.7 (5.7)	19.7 (26.8)	6.0 (4.9)	18.9 (9.0)
High Risk[b]	9.2 (9.1)	20.1 (15.1)	6.2 (5.6)	19.9 (26.1)	6.3 (4.0)	19.6 (6.6)
1999 LTCQ Incidence Measure						
Unadjusted[c]	8.4 (9.0)	19.9 (13.4)	6.9 (5.8)	20.2 (25.6)	6.0 (4.7)	17.9 (8.5)
Adjusted[d]	8.3 (8.6)	20.3 (12.5)	6.9 (5.8)	20.1 (25.8)	6.0 (4.3)	18.7 (6.3)

Notes:
[a] Low risk = All residents excluding high risk at most recent quarterly assessment
[b] High Risk = Impaired transfer or bed mobility, coma, malnutrition or end stage disease at most recent quarterly assessment.
[c] Unadjusted = Total # of residents with a higher pressure score than on previous assessment dived by the total number of residents with two valid quarterly assessments
[d] Adjusted = Facility mean PU incidence times the grand mean PU incidence, divided by the average predicted probability derived from a logistic model with transfer, unstable, history of PU, open lesions, wound care, bed rails used, bowel incontinence, bladder incontinence, bed mobility problems and locomotion problems as covariates.

The problem can also be examined by comparing the admission profile of residents in nursing facilities identified as the worst performers on the CHSRA and LTCQ pressure ulcer QIs. Table B.3 provides a comparison of the pressure ulcer prevalence rates at admission of facilities among the worst 10 percent and best 10 percent on the distribution for the two QIs.

Table B.3
Average Admission Pressure Ulcer Prevalence by Worst, Middle and Best Facilities as Determined by CHSRA and LTCQ QIs.

	Pressure Ulcer Prevalence at Admission					
	Worst 10% on QI		**Middle 80% on QI**		**Best 10% on QI**	
	Mean	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**
1999 CHSRA Prevalence Measure						
Low Risk[a]	.22	(.08)	.17	(.08)	.14	(.09)
High Risk[b]	.26	(.10)	.17	(.08)	.11	(.06)
1999 LTCQ Incidence Measure						
Unadjusted[c]	.23	(.13)	.17	(.08)	.10	(.07)
Adjusted[d]	.22	(.12)	.17	(.08)	.12	(.07)

Notes:

a Low risk = All residents excluding high risk at most recent quarterly assessment

b High Risk = Impaired transfer or bed mobility, coma, malnutrition or end stage disease at most recent quarterly assessment.

c Unadjusted = Total # of residents with a higher pressure score than on previous assessment dived by the total number of residents with two valid quarterly assessments

d Adjusted = Facility mean PU incidence times the grand mean PU incidence, divided by the average predicted probability derived from a logistic model with transfer, unstable, history of PU, open lesions, wound care, bed rails used, bowel incontinence, bladder incontinence, bed mobility problems and locomotion problems as covariates.

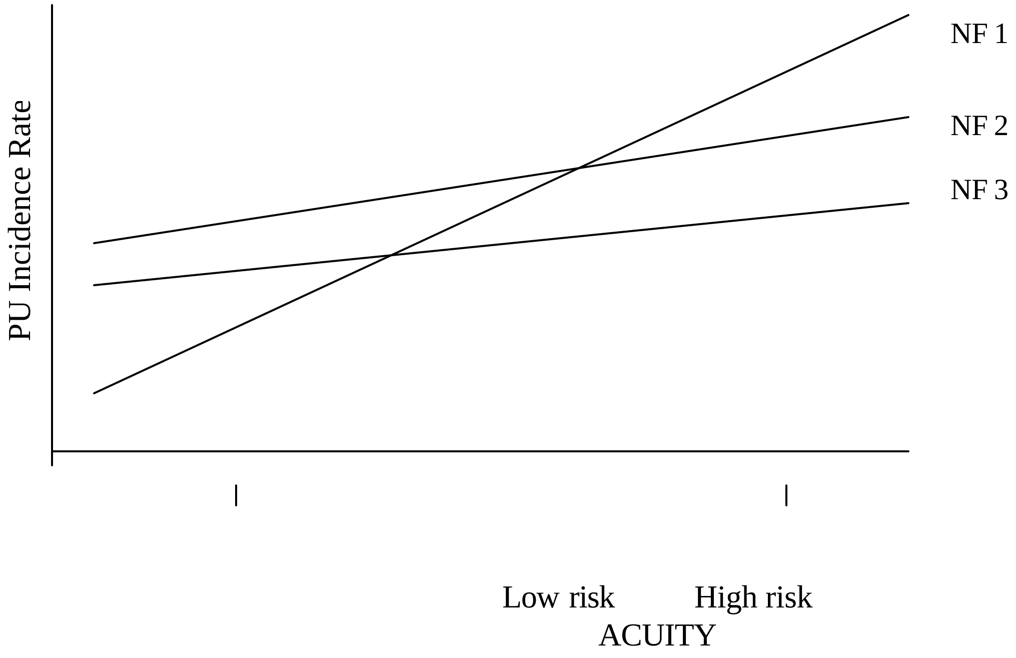
The above findings illustrate the shortcomings of existing Pressure Ulcer (PU) QIs. We propose to use multi-level modeling techniques to design a new PU QI to include a more robust risk-adjustment at the resident-level and a facility-level covariate to explicitly account for the selection and ascertainment bias problem.

There are many advantages to using a multi-level approach in developing QIs. The two most important strengths of multi-level modeling involve accounting for dependence among residents within a nursing facility and allowing the effects of covariates to vary from facility to facility.

Dependence among individual residents within the same facility can occur because treatment patterns or admitting decisions of a facility are similar for all residents. If we are unable to account for those facility-specific differences in our risk adjustment model explicitly, for instance because aspects of care are unobservable, our estimation may be imprecise or even biased. Multi-level modeling addresses this problem by incorporating a random effect for each nursing facility. The variability in these random effects is taken into account in estimating standard errors, thus reducing the likelihood of finding spuriously significant results. Simply speaking, by introducing a random effect for each facility, one acknowledges that facilities are different and one needs to account for that.

Relationships between individual resident risk factors and outcomes may vary from nursing facility to nursing facility. Some facilities may do a relatively good job treating low risk residents but do a comparatively poor job in their treatment of high-risk residents. Multi-level models allow the researcher to distinguish between such facilities in a parsimonious way. Figure 1 illustrates this phenomenon.

Figure B.1
Randomly Varying Slopes



Our use of multi-level modeling is designed to produce three main outcomes of the analysis:

1. How many facilities can we identify, with 90 percent confidence, as being in the best and worst 10 percent of facilities in terms of PU incidence?
2. How many facilities can we identify, with 90 percent confidence, as exceeding absolute benchmarks of 2 percent incidence for low-risk residents and 8 percent incidence for high-risk residents?
3. Can we demonstrate the value of using a facility-level variable to remedy the selection and ascertainment bias problems?

In the presentation of results below, the first of these approaches is used, although the results are similar using the other two approaches. In addition, our analysis includes several attempts to cross-validate our multi-level QI. These involve comparing how the multi-level estimates perform relative to the CHSRA and LTCQ QIs and other facility characteristics such as skilled staffing levels and examining the consistency of classifying facilities as "good" or "poor" over time using a subset of facilities.

Data Sources

Our analysis is based on all possible quarter-to-quarter incidence episodes in the first three quarters of 1999 for 564 nursing facilities in New York (159,041 cases, with the possibility of individuals appearing more than once). The resident-level outcome variable and covariates are computed from MDS quarterly assessments. The facility-level predictor of admission PU prevalence is based on admission assessments, which in 1999 were performed within the first five to seven days of admission for most residents.

Procedures for Multi-level Analysis

We used the statistical software program HLM 5.0 to fit a two-level logistic model with one resident-level covariate and one facility-level covariate. However, the standard errors of HLM estimates can be used to produce confidence intervals only if one makes the often unjustified assumption of normal distribution of the data. If this assumption is violated then estimates of the statistical precision may be incorrect. To generate reliable intervals around the predicted probabilities, it was necessary to adopt a hierarchical Bayesian approach to fitting the model using the software WinBUGS.

To facilitate convergence in the WinBUGS analysis, a single resident-level covariate was created as a composite of four risk factors (*bed mobility problems*, *incontinence*, *diabetes* and *hip fracture in last 180 days*). Based upon a variety of analyses, each risk factor was weighted to contribute a different amount to the composite measure based on its relationship to the outcome variable. For example, being dependent in bed mobility was about twice as important as diabetes (.95 vs. .42). In our random effect logistic model, each nursing facility has its own intercept and its own slope, for the resident-level covariate. Each facility's admission PU prevalence rate was included as our facility-level predictor.

For each facility, probabilities of pressure ulcer were calculated in two ways. The raw probability is the inverse logit of the linear predictor (i.e., of the sum of the intercept, resident-level term and facility-level term). For the standardized probability, the facility-level term is replaced by its mean value over all facilities before taking the inverse logit. The latter probability represents how a given facility would be expected to perform if it had an average admission pressure ulcer prevalence. Due to the non-linearity of the inverse logit, the magnitude of this adjustment varies among residents within each facility.

The model provides estimates of the variances of the random effects. We used the resulting facility-specific coefficients to generate profiles of high- and low-risk resident types to illustrate, for each facility, the average predicted probability that a resident with high or low risk would develop a pressure ulcer. The Bayesian analysis yielded facility-specific confidence intervals around each facility's predicted probability of producing a pressure ulcer when caring for high- and low-risk residents. We then transformed the estimates to examine how each facility performs against a given benchmark when treating high and low-risk residents. We chose two percent for low-risk and eight percent for high-risk residents.

Results

We first replicated the analyses presented above on the relationship between facility performance on the QI and the level of staffing in the facility. As can be seen in Table B.4, both the raw and the standardized multi-level pressure ulcer incidence measures are no longer inversely related to staffing levels. This suggests that this new approach reduces the previously identified selection bias differential substantially.

Table B.4
Relationship Between Staffing Levels and Worst/Best Performing Facilities Based on Multi-level High Risk QI

	Staff FTE Per 100 Residents					
	Worst 10% on QI		Middle 80% on QI		Best 10% on QI	
	RN Mean (SD)	RN+LPN Mean (SD)	RN Mean (SD)	RN+LPN Mean (SD)	RN Mean (SD)	RN+LPN Mean (SD)
1999 Multi-level Incidence Measure						
Raw	7.3 (6.7)	17.3 (10.8)	6.9 (6.1)	20.3 (25.5)	6.4 (3.7)	19.2 (6.2)
Standardized	6.4 (6.4)	17.5 (11.0)	7.0 (6.1)	20.3 (25.5)	6.5 (4.0)	19.2 (6.0)

Table B.5 clearly reveals that the standardized multi-level model incidence measure results in there being very little relationship between the prevalence of pressure ulcer incidence in the cohort of all residents admitted to the facility and whether the facility is in the “worst” or the “best” group of facilities with respect to pressure ulcer incidence in the long stay population. These results are in marked contrast to the strong linear relationship between the prevalence of pressure ulcers on admission and the CHSRA or LTCQ QI ranking based upon the long stay population. It should be noted that most of the admission cohort on whom the pressure ulcer prevalence on admission is calculated is **not** included in the calculation of the pressure ulcer incidence measure.

Table B.5
Multi-level Model: Average Admission Pressure Ulcer Prevalence by Worst, Middle and Best Facilities as Determined When Treating High Risk Residents

	Pressure Ulcer Prevalence at Admission					
	Worst 10% on QI		Middle 80% on QI		Best 10% on QI	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
1999 Multi-level Model Incidence Measure						
Raw	.22	(.11)	.17	(.08)	.11	(.06)
Standardized	.16	(.05)	.17	(.08)	.19	(.09)

Tables B.6 and B.7 present a cross-tabulation of the Multi-level model classifications and those emanating from the LTCQ and the CHSRA QIs that were recommended for general use in this report. As can be seen, there is considerable overlap between the methods, in that there is little or no shift from the "worst" to the "best", although considerable movement from the large group of facilities in the middle to the "top" or to the "bottom". Whether this is actually valid remains to be seen during the field validation effort still to be undertaken as part of this contract.

In comparing the multi-level model approach to the results of the CHSRA High Risk Pressure Ulcer QI in one case a facility was rated in the worst 10 percent using the multi-level approach and in the top 10 percent using the CHSRA model. Otherwise, about one-fifth of facilities ranked worst on the multi-level measure were also worst on the CHSRA but just under a third of facilities identified as best in the multi-level model were also in the top 10 percent on the CHSRA model. This is somewhat lower congruence than was observed between the multi-level model and the LTCQ model.

Table B.6
Cross-tab of High-Risk Multi-level QI Against LTCQ Adjusted Incidence Measure

	LTCQ Adjusted Incidence Measure					
1999 Multi-level Model	Worst 10%		Middle 80%		Best 10%	
Standardized Incidence Measure	Row %	N	Row %	N	Row %	N
Worst 10 %	39%	19	61%	30	--	
Middle 10 %	8%	37	84%	392	8%	38
Best 10 %	–		62%	29	38%	18

Table B.7
Cross-tab of Multi-level High-Risk QI Against CHSRA High Risk Prevalence Measure

	CHSRA High Risk Prevalence Measure					
1999 Multi-level Model	**Worst 10%**		**Middle 80%**		**Best 10%**	
Standardized Incidence Measure	**Row %**	**N**	**Row %**	**N**	**Row %**	**N**
Worst 10 %	20%	10	78%	38	2%	1
Middle 10 %	10%	43	82%	359	8%	38
Best 10 %	–	--	70%	33	30%	14

The final set of statistical validation analyses that were undertaken involved comparing the classification of homes as in the best or worst 10 percent of facilities in 1996 and in 1999. This is one of the more rigorous tests of the statistical validity of the QIs since it asks whether facilities that used to perform well or badly still do perform in this manner. In many respects, while one would like to see improvements in facility quality, there is an expectation that quality performance takes a while to change, meaning that one expects some degree of continuity, particularly among “good” facilities.

The analyses were performed based upon the 505 facilities (out of 536) that were operating in 1996 and in 1999 that met analytic sample requirements (sufficient numbers of observations in the denominator and numerator) and for which there were different facility survey data available from the two points in town. A separate multi-level model using the same variables and approach that was described above was constructed for 1996 and for 1999. In other words, in both 1996 and 1999 separate analyses were performed to classify facilities as among the “best” 10 percent or “worst” 10 percent performing in terms of the pressure ulcer outcome variable of interest. The resulting classification was separately done for the profile of “high” and “low” risk residents.

We found that the rank order correlation between the 1996 and the 1999 probability rates of having a pressure ulcer was .19 for the low risk profile and .25 for the high risk profile. Another way to consider the degree of congruence is to note that 16 of the “best” 50 facilities in 1999 were also classified as being “best” in serving high-risk residents in 1996. For the low risk facilities this level of congruence was 10 out of 50. Thus, among high risk cases, about one-third of facilities that were tops in 1999 had been tops in 1996 and among low risk cases this is true of one-fifth of facilities. Comparing these temporally determined analyses results with those of the multi-level results and the contemporaneously calculated CHSRA or LTCQ QI results, we find that the overlap in the distributions is quite similar.

Summary

The purpose of this technical appendix is to review the major conceptual and technical complexities associated with the creation and application of QIs for accountability purposes, that is for an audience outside of the facility itself. It is important to understand that conceptually any QI derived in the long-term care setting does not necessarily characterize the overall quality performance of a facility. This means that generalizing from a single measure is likely to result in an incorrect interpretation, for regulators, purchasers and consumers. Nonetheless, if these measures are to be used for accountability, they should have technical qualities that make them fair and representative of the relative performance of the facility in a domain of quality both now and in the near future, all other things being equal.

With these goals in mind, we used the example of pressure ulcer incidence and found that both of the measures that this report recommends for general use are lacking. We found that facilities with higher staffing had higher rates of pressure ulcers and that facilities admitting a higher proportion of residents with pressure ulcers had rates of pressure ulcer incidence in the long stay population 9 months later. Both of these findings suggest that the existing QIs are not "fair" since it is obvious that facilities admitting more clinically compromised residents with pre-existing pressure ulcers were being identified as performing worse on this QI. We all recognize that this phenomenon of selection bias occurs in the hospital sector, these results confirm that it operates in the long-term care sector as well. Furthermore, it is likely that ascertainment bias, the observation of differences in resident clinical problem rates due to more careful assessment practices and not due to real differences in the clinical picture, is intertwined with selection bias since facilities that admit more residents with selected clinical problems are likely to look for them more carefully.

In light of these related conceptual and technical deficiencies associated with existing recommended QIs, we developed an alternative approach using multi-level modeling and hierarchical Bayesian statistics to identify facilities that appeared to perform better or worse than expected in light of the mix of residents they serve and the pool of residents that they admit to the facility. We provisionally tested this model using MDS data from New York state for the period 1999. To provide some test of the stability of the resulting facility classification, we also used New York data from 1996.

Using this approach, we observed significant differences between the mean pressure ulcer incidence rates in New York (the intercepts). There were also significant differences from facility to facility in terms of the effects of the resident-level risk factor composite on the probability of developing a pressure ulcer (the slopes). The findings lead us to conclude with 90 percent confidence that a real difference exists between the facility with the lowest likelihood of producing a pressure ulcer for a resident with selected high risk or low risk characteristics and the facility with the highest likelihood of producing a pressure ulcer for the same type of resident.

Furthermore, once we had classified facilities as in the top 10 percent and the bottom 10 percent with respect to being significantly different from one another, we compared these facilities to both the staffing levels and the rate of pressure ulcer among residents admitted to the facilities. In both instances, we no longer observe evidence of selection bias since there was no relationship between pressure ulcer admission rates and the classification of homes in terms of the incidence of pressure ulcers in the long stay population.

Finally, we provisionally compared the classification of homes as being high or low in pressure ulcer rates in 1996 and 1999 and found some correspondence, about as much as there was between the other QIs and the multi-level approach. There are numerous reasons to not expect strong stability over that long a time period, but it was reassuring to know that there was a significant relationship between the two rates.

Implications and Discussion

The preliminary results of the analyses prompted by the proposed approach incorporated in this chapter are reasonably promising. However, they are not yet ready for general application. Although we have found numerous limitations in the existing QIs that we are recommending for general use, based upon our review of the general QI literature, their shortcomings are no worse (and indeed are likely to be less of a problem) than is the case for those in general use in the hospital, ambulatory or managed care organization contexts. Thus, we continue to believe that the careful application of the recommended QIs is important and may have a beneficial overall effect on the nursing facility industry, particularly as applied to the long stay resident population. There will be cases in which facilities that select high-risk residents for treatment will be scrutinized more carefully or unjustifiably sanctioned for these selection practices. This is unfortunate and mechanisms to minimize this occurring should be instituted in the survey and certification process.

We do believe that the multi-level approach that has been described in this chapter shows promise for broader application and overcomes some of the more egregious examples of unfair punishment for selection practices rather than quality performance. We propose to more formally test this approach in the field validation and subsequent analyses segments of the current contract.