Evaluation of the Quality Indicator Survey (QIS)
Contract #500-00-0032, TO#7
Final Report (December 2007)

Executive Summary

The QIS is a revised long-term care survey process that was developed under the Centers for Medicare & Medicaid Services (CMS) oversight through a multi-year contract. It represents an effort to standardize how the survey process measures nursing home compliance with federal standards and the interpretative guidelines that define those standards. The QIS demonstration represents the culmination of 15 years of CMS-sponsored research and development aimed at addressing criticisms of the long-term care survey process raised by consumers, providers, the General Accounting Office, Congress, survey agencies, and CMS Central Office.

The QIS is a two-staged survey process that was designed to produce a standardized resident-centered, outcome-oriented quality review. It uses an automated process that guides surveyors through a structured investigation intended to allow surveyors to systematically and objectively review all regulatory areas and subsequently focus on selected areas for further review. These features represent a departure from the Standard survey process and are consistent with measurement principles deigned to improve how consistently different surveyors conduct investigations. While the survey process has been revised under the QIS, the federal regulations and interpretive guidance remain unchanged.

In the fall of 2005, CMS launched a demonstration of the QIS through surveys of record by trained state survey staff. Five states were selected to participate in the demonstration: California, Connecticut, Kansas, Louisiana, and Ohio. The evaluation of the QIS was conducted in two phases. Phase I of the evaluation, the formative evaluation, focused on ways to improve the QIS, particularly in the domains of time and efficiency. The second, summative phase of the evaluation focused on how well the QIS achieved its primary objectives of improving the accuracy, consistency, and documentation of the nursing home survey process within existing survey resources, addressing these research questions:

- ***Does the QIS lead to increased accuracy?*** The accuracy of the QIS was assessed through site visits to 20 nursing homes. While on-site, members of the research team collected data on a series of quality indicators based on protocols that were developed for several domains of nursing home care. These quality indicators give measures of the quality of care provided by the nursing home that can be compared to surveyor findings to determine whether there is more agreement between surveyor findings and the quality indicators for QIS or Standard surveys. If the QIS is more accurate than the Standard survey, then we expect that there would be a higher level of agreement between QIS survey results and the quality indicators than for Standard surveys.

- ***Does the QIS result in improved documentation of survey deficiencies?*** The question of whether the QIS leads to improved documentation was addressed using the CMS 2567 Forms generated by surveyors. We used content analysis to compare the quality of documentation review for a sample of 2567 Forms from QIS and Standard surveys, using a blind review process to measure whether the documentation supports the specific F-tag, scope and severity that was cited.

- ***How does the time required to complete the QIS compare to the time required for the current survey?*** A key research question is whether the QIS requires more surveyor time than the current survey process. Using data from the CMS-670 form, we analyzed how QIS time compared to the time for the facility's prior survey and to Standard surveys at similar facilities. This analysis also examined factors that are related to survey time requirements such as facility size and the number of deficiencies cited by surveyors.

- ***How does the QIS impact the number and types of deficiencies that are cited?*** Analysis of survey deficiencies was used to examine the impact of the QIS on the number and scope/severity of deficiencies cited and also whether the QIS is associated with changes in the types of regulatory care issues that are cited. We examined survey deficiency results separately for each state and used a combination of pre-post and cross-sectional analyses to explore the impact of the QIS on survey deficiencies.

- ***Does the QIS improve surveyor efficiency?*** One of the objectives of the QIS is to improve the efficiency of surveyors by focusing survey resources on facilities that have the largest number of quality concerns. We analyze the relationship between time and outcomes to measure whether the QIS is associated with changes in how well surveyor time is targeted to facilities with more quality problems

This is an evaluation of real-world implementation of the QIS with several constraints on ideal research design. The study design focused on controlling two factors that are known to impact survey results: state differences and changes over time in survey practices. It was not possible to control for the selection of surveyors or the facilities receiving a QIS survey, and the design was also constrained by limited funding and time to complete the evaluation. In general, comparisons focus on within state differences between the QIS and Standard surveys.

Does the QIS Lead to Increased Accuracy?

Improved accuracy of quality of care and quality of life problem identification is one of the objectives of the QIS. The QIS includes substantially larger random samples of residents including 40 residents currently residing in the facility and up to 30 admissions from the prior six months. This is expected to yield more valid inferences about the care provided to residents and systems of care. Improved accuracy may also result from the Stage Investigative protocols, standardized data collection, and use of computerized algorithms to identify areas on which to focus in Stage II.

To test whether the QIS is more accurate than the Standard survey, we made site visits to a sample of 20 facilities to collect data on the quality of care provided by the facility. The sample size was determined by the time and resources available for this activity. While data were collected at the same time as the nursing home's survey, the evaluation researchers worked independently of the surveyors, collecting information on a series of Care Indicators (CIs) to assess quality in five domains: incontinence, nutrition, pressure ulcers, choice, and activities. Differences between the QIS and standard survey in how accurately they measure the care elements described in the interpretative guidelines can only be proven if the deficiencies written by the two survey types are compared to deficiencies that might be written by an independent assessment of care quality. A more accurate process would be expected to identify more of the same care problems identified by the independent system. We used information collected as part of the site visits to address a set of research questions related to the ability of the QIS to detect quality problems compared to the Standard survey.

Methods

Our approach for investigating whether the QIS is more accurate than the Standard survey was to compare the relationship between facility quality as measured by the CIs and survey findings for QIS and Standard surveys. If the QIS is more accurate than the Standard survey, then there should be a stronger relationship between quality and deficiencies for QIS surveys than for Standard surveys. Overall, there were 75 CIs that were included in our analyses. These included some that were previously tested and validated in the Assessing Care of Vulnerable Elders (ACOVE) project as well as some new indicators were combined to assess quality in five care areas: incontinence, nutrition, pressure ulcers, choice, and activities. The CIs that were used could be linked with specific care processes that were identified for investigation in the interpretative guidelines or critical element pathways for the specific care areas selected for investigation.

We followed very specifically defined protocols for observation and data collection, mapping CI scores to CMS guidelines for compliance, and scope and severity decisions using a set of uniform decision rules. All of the CIs use an "if-then" framework that describes the residents to whom the CI applies and the care process to be provided. A set of specific criteria were used to identify residents at risk for various quality problems, and the indicators were only scored for residents who met these criteria. The scoring system produced continuous data expressed as the number or percentage of residents that failed each CI.

The translation of CI data into Ftag statements was particularly problematic for this project since the current rules used by CMS allow considerable judgment on the part of surveyors as to when survey results raise to the level of a citation. We preferred not to depend on the expert judgment of research staff in linking CI results to probable Ftag citations and instead developed standardized rules which we believed to be consistent with the intent of the CMS regulations. The basic rule that we developed was that a resident must fail an indicator in assessment, care planning or care plan revision, and provision of care to site an Ftag. Ftags were cited as greater than isolated if more than two residents met the above criteria in specific areas of care.

The data collected on-site were used to compare the accuracy of QIS and Standard surveys. To do this, the data were aggregated across CIs into the five care areas and compared to related Ftags that were cited by the state surveys that were conducted concurrently with our site visit. CMS 2567 forms were carefully read by two independent reviewers to identify Ftags that were related only to the targeted five care areas; we did not consider Ftags that were not related to the five care areas. The expectation was that a high failure rate for the CIs in a domain would correspond to a failure to meet regulatory standards, which should be related to the Ftags cited by surveyors. Our test for whether the QIS is more accurate than the Standard survey is whether there was a stronger relationship between the CIs and survey findings for QIS surveys than for Standard surveys.

Site Visits

Site visits were made to two matched pairs (one QIS survey, one Standard survey) in each state for a total of 20 site visits. We considered a number of criteria in selecting the matched pairs, including survey region, facility size, ownership type, survey deficiency history, and recommendations from State Survey Agency contacts. It was not possible to match on all of these criteria, but we selected pairs of facilities that appeared to be a good match based on an overall assessment of our matching criteria. We did not visit nursing homes that were undergoing

a Federal Monitoring and Oversight Survey or nursing homes that State Survey Agencies asked us not to visit. Two-person teams consisting of at least one registered nurse were used on all site visits. With the exception of one member of the team, all of the site visitors had collected data for the Formative Evaluation and were thus familiar with the project. While on-site, researchers collected information needed to calculate pass/fail rates for each of the CIs. There were three components of the data collection process: medical record review, resident interview, and observation.

- ***Medical record review:*** Medical record review included a screening tool to identify residents who had or were at risk for specified conditions, a review of diagnoses and medications that may impact a resident's condition, and a chart review investigation.

- ***Resident interviews:*** All non-cognitively impaired residents in our sample were approached for an interview, which included questions about pressure ulcers, urinary incontinence, nutrition, choices, and activities.

- ***Resident Observations:*** Four types of resident observations were completed during the onsite visit: Continuous observations for Toileting and Positioning, 60-minute Behavioral Observations, Dining Observations, and ADL-Choice observations.

Information was collected from a sample of residents who had been in the nursing home for 12 months or less and who had or were at risk for one or more of the domains of interest. The sample consisted of a minimum of ten residents with each condition, at risk for the condition, or having a history of the condition. In most cases, residents had more than one condition resulting in a sample size of between 12 and 20 residents per site.

Results

In general, we did not find any differences in accuracy for the QIS and Standard surveys.

- ***The QIS and Standard survey samples were comparable with respect to overall quality and survey deficiencies cited.*** The two groups were also similar with respect to the frequency of Ftag citations of a scope beyond isolated. These findings suggest that the matching criteria used to select facilities for the site visits worked reasonably well.

- ***The overall failure rate on the CIs was high.*** The overall fail rate across all care areas was 44 percent for the Standard sample and 45 percent for the QIS sample. The high CI fail rate in both QIS and Standard survey facilities supports the argument that many Ftags could have been written in most facilities even if rigorous standards were used to convert quality findings into deficiency statements. Across nursing homes, there was significant variability between CI fail rates for all care areas that provided the opportunity to judge how the two survey types discriminated quality with the Ftag citation system.

- ***Overall, the relationship between quality and survey deficiencies was low.*** We found a positive but low correlation between the overall number of related Ftags and quality as measured by the CI fail rate for both Standard and QIS surveyed nursing homes. Neither QIS nor Standard surveys consistently documented that providers failed to implement many of the care indicators recommended in the investigative protocols. Some recommended quality measures were never or rarely documented by either survey team in an Ftag statement despite the fact that the majority of residents who were eligible for the care described in the guidelines were found in this evaluation to not receive that care.

- ***There was no evidence of a stronger relationship between quality and deficiencies for QIS surveys.*** We found that there were no differences in the ability of the QIS and Standard surveys to detect quality problems as measured by the explicit quality assessment protocols used by the research teams. The correlation between Ftag citation rate and overall indicator fail rate across all care areas was positive but relatively low for both types of surveys. It was higher for Standard surveys (0.34) than for QIS surveys (0.19).

- ***Findings suggested that more survey deficiencies with scope greater than isolated could have been cited for both QIS and Standard surveys.*** We found that over fifty percent of the facilities could have been cited for a scope greater than isolated in all care areas, with the exception of activities, even when a rigorous standard was used to translate the CI data into deficiency statements. Both types of surveys failed to detect more than isolated problems in many facilities.

- ***There was no evidence that the QIS was more accurate with respect to survey deficiencies with scope greater than isolated.*** The CIs identified 29 occasions that would justify an Ftag greater than isolated across the Standard survey nursing homes and 31 opportunities for QIS nursing homes. The Standard survey teams wrote greater than isolated Ftags in 15 out of these 29 opportunities (52 percent) vs. 11 out of the 31 opportunities (35 percent) for QIS surveys.

- ***Both types of surveys failed to detect many residents with poor pressure ulcer and weight loss outcomes.*** We developed a set of rules for converting the CI data into citations that could have been made beyond G for weight loss and pressure ulcers. We identified seven different QIS nursing homes and five different Standard nursing homes could have been cited for a G or higher pressure ulcer citation. All ten QIS nursing homes and eight Standard nursing homes could have been cited for a G or higher citation in nutrition. Both types of surveys failed to detect many residents who have poor pressure ulcer and weight loss outcomes and who also receive poor care according to multiple data sources.

- ***The QIS may be better at citing appropriate G-level deficiencies, but the available sample size is insufficient to draw any conclusions.*** There were a total of two Ftags cited at the G level for nutrition and two cited at G-level or above for pressure ulcers. All the citations were by QIS teams in one state and took place in just two different nursing homes. Given the small sample size, it is not possible to draw any conclusions from these findings, but they suggest that the QIS may be better at citing appropriate G-level citations.

Implications

We did not find evidence that the QIS is more accurate than the Standard survey, despite the fact that it has started the process of making the survey process more specific and focused with its Stage I protocols and automated data entry system. We qualify these findings by noting that comparisons between the QIS and Standard surveys were limited by a small sample size; thus the data we provide are best used for survey improvement purposes rather than to inform a decision about what type of survey process to use. That said, we do not believe that a larger sample size would produce dramatically different results until further refinements are made in the basic concepts that underlie the QIS and which make it different from the Standard survey process. Based on the data collected in the formative evaluation and this field test evaluation, we believe that the best explanation for the lack of differences between the two survey methods is related to two issues: 1) the specificity of the investigative guidelines and the critical element pathways,

and 2) how feasible or "user friendly" the critical element pathways and interpretative guidelines are to implement. The various investigative documents used by both survey types vary in specificity such that there is much interpretation left to the discretionary judgment of surveyors. Despite this evaluation's small sample size, the recommendations for correcting these specificity and feasibility problems are clear.

Does the QIS Result in Improved Documentation of Survey Deficiencies?

Because of the larger samples, more structured data collection, and the automated computer processes used to organize survey findings, the QIS was expected to lead to improved documentation quality. The Critical Element Pathways, intended to assist surveyors in their Stage II information gathering, were expected to lead to the citation of more related deficiencies.

Methods

To investigate whether the QIS results in improved documentation of survey deficiencies, we conducted a content review and analysis of a sample of draft statements of deficiencies from CMS 2567 forms. Pairs of survey citations (one from a QIS survey and one from a Standard survey) were reviewed and scored by trained nurse reviewers using a set of review protocols that were developed based on the CMS Principles of Documentation, the State Operations Manual, and input from CMS survey and certification staff at both the Central Office and Regional Offices. The protocol was intended to capture four separate dimensions of survey deficiency documentation, including:

- Quality of the evidence that justifies or substantiates the citation
- Accuracy of the scope rating
- Accuracy of the severity rating
- The extent to which surveyors issue the related process tags when an outcome tag is cited

Rather than select *complete* 2567 forms for review, we selected specific Ftags from the draft forms, in order to analyze matched pairs of Ftags cited on standard and QIS surveys. The sampling strategy was driven to a large extent by the available sample of Ftags for which there were pairs of deficiencies within State (for QIS and Standard surveys) at similar levels of scope and severity.

In order to capture surveyor findings as close to the time that impressions about facility quality and compliance with federal LTC requirements were formulated by the surveyors, we chose to use the "draft" 2567s, or initial reports of survey findings, in this analysis. These are survey findings that have not yet undergone supervisory review. The decision to review draft 2567s was the result of careful consideration by the research team and CMS of how best to measure the difference between the standard and QIS documentation quality. The QIS was not designed to impact the supervisory/office review process, and it was decided that a better test of whether the QIS leads to improved survey documentation would use the version of the 2567 that is produced by the survey team, before any supervisory/office review.

We requested that each State send us draft 2567s on an ongoing basis for both standard and QIS surveys. We trained a team of two researchers who reviewed a sample of the 2567s using standardized review protocols, and scored the 2567s and associated Ftags according to the quality of documentation. The Ftags were reviewed blind as to whether they were from a QIS or Standard survey.

These data were analyzed to determine whether the QIS does in fact produce better documentation than the Standard survey process. Two major limitations of the analysis are the lack of formal tests of inter-rater reliability and the fact that each Ftag was reviewed by only one reviewer. These limitations were driven by the time and resources available for this activity. The procedures for establishing inter-rate reliability between the reviewers were really informal practice sessions on a handful of tags, followed by discussions to resolve differences and subsequent changes to the protocols and guidance.

130 Ftags selected from the top 15 deficiencies in four demonstration states were rated for documentation adequacy by two reviewers using a research protocol. (Kansas was excluded for technical reasons.) Within each deficiency area, the tags were selected from QIS and Standard surveys in equal numbers and roughly the same scope and severity. The reviewers were both nurses formerly employed as long-term care surveyors. The review protocol had four dimensions: quality of the evidence that justifies a citation; accuracy of the scope rating; accuracy of the severity rating; and the extent to which surveyors issue the related process tags when an outcome tag is cited.

Results

Although there were some differences across the dimensions examined, overall there was essentially no evidence that the QIS leads to higher quality deficiency documentation. Nor was there any evidence that the QIS led to an overall increase in the citation of Related Ftags:

- ***Standard surveys were more likely to include both a deficiency statement and a related outcome.*** Overall, 33 percent of Standard survey Ftags reviewed was noted to include both a deficient statement and a related outcome, while 21 percent of QIS Ftags reviewed met this standard. QIS survey *process* deficiencies, such as assessment (F272) and care planning (F279) deficiencies, were more frequently accompanied by their related outcome tags than were Standard survey deficiencies, but, for outcome deficiencies (e.g., pressure ulcer development, F314), Standard surveys were more frequently accompanied by related process deficiencies than were QIS outcome deficiencies.

- ***QIS deficiencies tended to cite more types of evidence than Standard deficiencies.*** QIS Ftags reviewed cited as many or more types of evidence in general, for both process and outcome tags.

- ***There was little difference with respect to the quantity of evidence cited***. There were differences across States. California and Ohio QIS surveys referenced a higher number of data points than their Standard survey counterparts, while fewer data points were cited on Louisiana QIS surveys.

- ***There was no evidence that the QIS was associated with citation of additional related Ftags.*** The review of CE Pathways did not reveal significant differences between the "related Ftags" cited on QIS vs. Standard surveys. It does appear that the ADL Critical Element Pathway consistently guided QIS surveyors to cite a higher average number of related deficiencies; however, no other distinct pattern emerged from this review to support that the availability of the CE Pathways in the QIS survey influenced deficiency citations.

Implications

Given the initial difficulty in achieving inter-rater reliability on this aspect of the 2567 review, and the limited sample of reviewed tags in this component of the review, the CE Pathway review findings should be interpreted with caution. While there is no concrete evidence that a reasonable level of inter-rater reliability was achieved prior to the review nor sustained during the review, the reviewers were experienced surveyors who participated in the development of the review protocol and guidance. The lack of a systematic differences in documentation quality may reflect the variable knowledge and skill of the surveyors under both the QIS and Standard survey, which likely influences both the decision to cite and the supporting documentation

How Does the Time Required to Complete the QIS Compare to the Time Required for the Current Survey?

A major evaluation question is whether the QIS takes longer than the current survey process. The formative evaluation was used to identify a number of ways to streamline the QIS to reduce survey completion times. For national implementation, it is important that the QIS be resource neutral, at least in the aggregate across states. Increased time requirements may be problematic given limited survey budgets.

Methods

The data source for these analyses is Form CMS-670, the Survey Team Composition and Workload form. Surveyors use this form to record the amount of time that members of the survey team spend on pre-survey preparation, on-site, travel, and off-site report preparation. For these analyses, we considered pre-survey preparation, on-site, and off-site report preparation time, but not travel hours. Our analyses emphasize a time measure that includes only the time required to complete the survey itself and that does not include time associated with post-survey follow-up activities such as ensuring that facility plans of correction have been implemented. We believe that this allows for the most accurate analysis of the time associated with QIS surveys. However, because the time associated with survey follow-up activities may be indirectly related to the QIS (e.g., if this follow-up time is related to the number of deficiencies and the number of deficiencies is higher for QIS surveys), for some analyses, we used a time specification that also included the time associated with post-survey followup activities.

We estimated the time requirements for the QIS by comparing the change in survey completion times for QIS surveys and Standard surveys conducted during the same time period to the previous survey at the facility, which was completed using the Standard survey process at all facilities. We used a “difference in differences” model in which we compared the change in survey completion times for facilities that had a QIS survey to the change in time for facilities that had a Standard survey during the QIS demonstration. This analysis allowed us to take account of factors, other than the QIS, that affected survey completion times that would be missed in a simple pre-post comparison.

Results

- ***Results varied across States.*** While we used a variety of analyses to examine the time requirements of the QIS, in most cases, the most appropriate analysis was a comparison of survey completion times for the QIS and the prior survey, which was completed using the Standard survey process. The pre-post comparison showed that results varied across States.

- For three states (California, Kansas, Ohio), the QIS took longer than the prior Standard survey at the same facilities—the pre-post differences were especially large in California and Kansas. In California, the QIS survey required 68 more hours than the prior survey at the facility, an increase of 46 percent. In Kansas, the QIS required 62 more hours than the prior survey, a 43 percent increase. In Ohio the difference was small; the QIS took an additional 4.55 (3.6 percent) hours, a difference that went away entirely for surveys completed after the changes recommended as part of the Formative Evaluation were implemented.

- In Connecticut and Louisiana, the QIS was completed more quickly than the prior survey at the facility. In Connecticut, the QIS took about nine hours less than the prior survey. In Louisiana, the difference was very large. The QIS took 46 hours less than the prior survey at the facility. There was a large decrease in time for both Standard surveys in the State that may reflect disruptions to the survey process resulting from Hurricane Katrina.

- Inclusion of the time associated for post-survey follow-up activities (e.g., to ensure compliance with plans of correction) did not change the basic conclusions of our analysis. Results varied across the five states, with QIS surveys taking much longer in California and Kansas, slightly longer in Ohio, and less time in Connecticut and Louisiana.

- ***Exclusion of outliers does not change basic conclusions regarding QIS completion time.*** There were some QIS surveys that took an extraordinarily long time to complete, in some cases 200 or more hours. All but one of these surveys was in California or Kansas. The explanation for the high completion times varied. Some of these surveys were among the first QIS surveys conducted in the state; for others there were extenuating circumstances that affected the time needed to complete the survey. Survey times were extremely high for QIS surveys completed in California's East Bay region, in which surveyors had difficulties becoming compliant with QIS processes. For other high time outliers, the time required to complete the QIS was not much higher than the time required to complete the prior survey, likely due to a large number of deficiencies cited on both surveys.

- ***The changes to the QIS implemented after the formative evaluation reduced QIS time requirements.*** The changes to the QIS implemented after the formative evaluation appeared to lead to modest reductions in the time required to complete QIS surveys.

  - In Connecticut, the mean time for QIS surveys was 126 hours for surveys started before May 31, 2006, compared to 105 hours for surveys started after implementation of the formative evaluation changes.

  - For Kansas, mean survey completion times were almost identical for the two periods, but median time was 20.5 hours lower for surveys that started after implementation of the formative evaluation recommendations.

  - There were only six QIS surveys conducted in Louisiana before May 31, but both the median and mean times were higher for these surveys than for surveys that were started on May 31 or later.

  - In Ohio, average survey completion time was 133.9 hours for the earlier group of surveys, compared to 125.6 hours for surveys that applied the changes made following the formative evaluation.

- Across all states, the average overall survey completion time was 133.9 hours for the earlier group of surveys, compared to 125.6 hours for surveys that reflected the changes made following the formative evaluation.

- Note that California did not begin doing QIS surveys until June 2006, which was after implementation of the formative evaluation changes.

- ***There was some evidence of a learning curve.*** We anticipated that, as they became familiar with the QIS process, surveyors would be able to complete the QIS more quickly We found some evidence of a learning curve in the two states (Connecticut and Ohio) that had stable survey teams throughout the demonstration period. These results suggest that the comparisons of QIS time that only include the subset of surveys completed after implementation of the formative evaluation recommendations likely gives the best estimates of the long-term impact of the QIS on survey completion times. A limitation of these analyses is that the effects of a potential learning curve are confounded with other changes that occurred during the demonstration period such as the changes to the QIS made after the formative evaluation.

Implications

The experiences of Connecticut and Ohio suggest that the there is nothing inherent about the QIS process that suggests that the QIS survey cannot be completed in the same amount of time as Standard surveys. In Connecticut, QIS surveys appeared to take less time than Standard surveys. In Ohio, QIS surveys completed after implementation of changes to the QIS following the formative evaluation took about the same time as the prior survey at the facility. Based on the experiences of these two states, we would conclude that concerns about the time required to complete the QIS survey are unfounded, as there is nothing to indicate that the QIS takes any longer than the Standard survey process.

The experiences of California and Kansas, however, lead to the opposite conclusion, suggesting that it is not automatically the case that the QIS can be completed as quickly as Standard surveys. In Kansas, QIS surveys took an average of 207 hours to complete, 62 hours more than the prior survey. The time differences that we observed in the state cannot be attributed to a learning curve, outliers, or other factors that may not affect long-run survey time requirements. Something about the way that QIS surveys were conducted in the State led to the QIS surveys taking much longer than Standard surveys. In California, QIS surveys also took considerably longer to complete than the prior survey. This was the case for both the San Diego and East Bay districts, reflecting the difficulties that the State experienced in conducting QIS surveys. Given our mixed findings with respect to QIS completion times, if the QIS is expanded to additional states, it is difficult to know whether these states' experiences will be more like those of Connecticut and Ohio or those of California and Kansas. As a result, while we do have conclusions about how long QIS surveys took to complete in each of the demonstration states, we do not offer conclusions about the time requirements of the QIS in other states. The likelihood is that there will be some states for which the QIS does not take any longer to complete than Standard surveys and others that struggle to implement the QIS and find that it does take longer. In any case, the results for Connecticut, Louisiana, and Ohio lead us to conclude that it is certainly possible for states to implement the QIS in a resource-neutral way, even taking account of any additional post-survey follow-up activities that may result from additional survey deficiencies cited on QIS surveys

It is important to keep in mind that, in order to be budget neutral, we do not require that the QIS take the same amount of time or less in every state. The experiences of the five demonstration states suggest that the QIS may be resource neutral in the aggregate. It may be that some reallocation of survey and certification resources would be required if the QIS were implemented in other states, with additional resources given to states for which the QIS takes longer than Standard surveys. It was beyond the scope of this evaluation to address these types of issues.

How Does the QIS Impact the Number and Types of Deficiencies That Are Cited?

While increases in survey deficiencies are not among the stated objectives of the QIS, we examined the impact of the QIS on the number, scope/severity, regulatory care areas, and individual F-tags cited by survey teams. We also examined whether the QIS leads to a more comprehensive assessment of all regulatory areas. We estimated the impact of the QIS on these outcomes using a combination of pre-post and difference-in-difference analyses.

Methods

We used data from the CMS Online Survey Certification and Reporting (OSCAR) system to measure survey outcomes and Form CMS-670 data to measure the impact of the QIS on survey outcomes. We examined the number of deficiencies, the number of deficiencies for actual harm and above (G-level and above, and the regulatory care areas cited by QIS surveys. Our data do not include complaint surveys and we considered only deficiencies that resulted from health inspections. As we did with the analyses of QIS completion time, we used a "differences-in-differences" model to estimate the impact of the QIS on survey outcomes. Using the difference-in-differences model, we can estimate the impact of the QIS on survey completion time for the other states as well, allowing us to adjust for factors that may have affected survey outcomes across both standard and QIS surveys.

Results

- ***The QIS was associated with an increase in the number of survey deficiencies:*** Using the difference-in-differences model to account for general time trends, we estimate that the QIS was associated with 1.6 additional deficiencies in California (a 14 percent increase), 0.6 fewer deficiencies in Connecticut (a 9 percent decrease), 9.4 additional deficiencies in Kansas (a 99 percent increase), 1.9 additional deficiencies in Louisiana (a 29 percent increase), and 2.4 additional deficiencies in Ohio (a 52 percent increase).

- ***The QIS was associated with an increase in G-level deficiencies:*** We also examined the impact of the QIS on deficiencies cited at the G-level or above (G, H, I, J, K, L). The rate of G-level deficiencies was relatively small for both types of surveys, but the QIS was associated with large increases in Kansas, and Ohio and a large decline in Connecticut. There was a slight increase in California and relatively little change in Louisiana.

- ***The QIS was associated with an increase in the regulatory care areas cited***: One of the objectives of the QIS is to comprehensively review a wide range of regulatory care areas. In all five states, there were more regulatory care areas cited on QIS surveys than on the prior survey at the facility. For all regulatory care areas except for infection, facilities were more likely to receive a deficiency with the QIS survey than with their prior survey. For some regulatory care areas (resident rights, quality of life, dietary care, physician services, dental care, physical environment), the differences were substantial.

Implications

The analysis provides strong support for the hypothesis that the QIS leads to an increase in the number of survey deficiencies and an increase in the regulatory care areas that surveyors cite, supporting expectations about the QIS. These are an important findings given the studies by the General Accounting Office (GAO) and Office of the Inspector General (OIG) that have found that the Standard survey under reports deficiencies, harm-level deficiencies, quality of life, resident rights, and dental deficiencies. As a practical matter it would be difficult to implement any system that results in several fold increases in deficiencies, but this was not in general the case although the increase observed for Kansas QIS surveys may be a reason for concern.

A potential limitation of this analysis is that we are unable to control for surveyor quality. QIS surveyors were chosen to participate in the demonstration because of their experience. It may be that the QIS teams have higher citation rates than other survey teams, and that this may explain some of the increase in survey deficiencies that we observed for QIS surveys.

Does the QIS Improve Surveyor Efficiency?

One of the objectives of the QIS is to improve the efficiency of surveyors by focusing survey resources on facilities that have the largest number of quality concerns. The two-staged process focuses surveyor resources on areas identified as problematic in Stage I and permits bypassing the second stage investigation for potential problems if these concerns do not exceed thresholds established by prior research. If the QIS is successful in achieving the objective of increased efficiency, then there should be a stronger relationship between survey time and survey outcomes for QIS surveys than for Standard surveys.

We examined the relationship between the time required to complete surveys and survey outcomes for QIS and Standard surveys. Further, we examined whether the QIS was associated with a change in this relationship that suggests that the QIS is more effective than the Standard survey in terms of focusing surveyor resources on the most problematic facilities.

Methods

To examine whether the QIS was associated with an increase in surveyor efficiency, we estimated a series of regression models that included interaction terms for the number of deficiencies and the type of survey (QIS, prior survey at QIS facilities, Standard survey during demonstration period).

Results

We found that, for both QIS and Standard surveys, there is a strong relationship between the total number of deficiencies and survey completion times. The overall correlation between time and deficiencies was 0.56 for Standard surveys (including both pre-QIS surveys and Standard surveys conducted during the demonstration period. Within individual states, this correlation ranged from 0.39 in Connecticut to 0.77 in Louisiana. For QIS surveys, the correlation was 0.73 across all QIS surveys. The within-state correlations were lower (ranging from 0.53 in California to 0.75 in Louisiana).

The regression models showed that number of deficiencies and facility size were strong predictors of survey completion time, while G-level deficiencies were typically not significantly related to survey completion time. Regression results showed that, while there was a strong relationship between time and deficiencies for QIS surveys in Connecticut, Louisiana, and Ohio, the only state in which the QIS was associated with an increase in surveyor efficiency (as measured by the relationship between time and survey outcomes) was Ohio. For Ohio, the coefficient on total deficiencies for QIS surveys was much higher than the coefficient on total deficiencies for pre-QIS and Standard surveys, suggesting that the QIS led to improved surveyor efficiency.

The limited number of QIS surveys available for the regression models in California, Kansas, and Louisiana is an important limitation of this analysis, as it restricted our ability to examine subsets of QIS surveys such as those completed after an initial learning process or those conducted after implementation of the changes to the QIS that followed the formative evaluation. The patterns that we observed in Ohio suggest that the QIS has the potential to improve surveyor targeting to facilities with the most quality problems, but the experiences of the other states suggests that this need not be the case.

Implications

These results do not appear to be consistent with the expectation for improved targeting and efficiency for the two-staged QIS process. The only state consistent with this expectation was Ohio, which had a lower correlation between Standard survey time and deficiencies than the other states and far more opportunity for improvement. The small sample size available for these analyses limits the conclusions that we can draw for California, Kansas, and Louisiana.

Conclusions

The results of the evaluation were mixed and do not lead to firm conclusions about the effectiveness
of the QIS.

- ***Does the QIS lead to increased accuracy?*** Based on the relationship between survey findings and a set of care indicators intended to measure the quality of care provided by nursing facilities, we did not find evidence that the QIS was more accurate than the Standard survey. Our results suggested that more survey deficiencies with scope greater than isolated could have been cited for both QIS and Standard surveys. Ultimately, under both types of surveys, there appears to be a great deal of surveyor discretion and judgment that influences the decision to cite.

- ***Does the QIS result in improved documentation of survey deficiencies?*** We found essentially no differences in documentation quality associated with the QIS, although interrater reliability concerns limit the strength of this conclusion.

- ***How does the time required to complete the QIS compare to the time required for the current survey?*** Results indicate that there is nothing inherent in the QIS which indicates that it cannot be resource neutral. We found that the QIS took considerably longer to complete than Standard surveys in two of the five demonstration states; two states consumed about the same amount of time and one state's time was open to different interpretations.

- ***How does the QIS impact the number and types of deficiencies that are cited?*** The results of this evaluation clearly indicate that the QIS cites more deficiencies, at higher levels, and more in these usually under-cited areas.

- ***Does the QIS improve surveyor efficiency?*** The correlation between time and deficiencies was higher for QIS surveys than for Standard surveys. Ohio was the only state for which the QIS was associated with an increase in surveyor efficiency. A number of recommendations for improving the QIS emerged from the field work conducted as part of the summative evaluation and the earlier formative evaluation. These recommendations focused on ways to improve the accuracy of the QIS.

- ***Improve specificity and usability of investigative guidelines.*** The care elements that are recommended for investigation in existing interpretative guidelines and critical element pathways should be modified so that they are consistent with the principles that guide reliable and accurate measurement.

- ***Provide competency-based training for surveyors to improve consistency.*** Provide survey staff with training in the principles of reliable measurement and document that the trained surveyors can use the investigative protocols to produce consistent and accurate quality conclusions.

- ***Evaluate how well the QIS Stage I and unstaged protocols identify problem areas that should be investigated in Stage II***. If the QIS is accurately detecting areas for investigation, then quality measures for facilities that are flagged for an investigation should be different and worse than the measures for facilities that are not flagged. We did not find these differences in our evaluation of QIS accuracy, suggesting that the question of whether Stage I accurately identifies areas in which there are potential quality problems and which are thus the best targets for Stage II investigations is relevant.

- ***Increased structure in Stage II to make decision making more explicit in determining noncompliance, scope, and severity.*** Despite the structure of Stage I that ensures that surveyors conduct a more comprehensive survey and utilizes more information from residents and families, the process becomes increasingly subjective in Stage II and during certain facility-level tasks. Of highest priority is the development of additional CE pathways for the many important care areas where these do not exist. A second priority is to improve the integration of the CE pathways into the Stage II investigation.

This report also includes a chapter "Quality Indicator Survey Demonstration: The Big Picture" that was written by Andrew Kramer, one of the developers of the QIS. This chapter summarizes findings from the University of Colorado's work in developing the QIS and recommendations for future development work. We requested that Dr. Kramer write this chapter because of the many insights that he has gained during the development of the QIS, but it is not part of the independent evaluation of the QIS conducted by Abt Associates and Vanderbilt University.