Costs and Effects of Participation in Collaboratives to Improve Chronic Illness Care

Technical Appendix

EMMETT B. KEELER & GEOFFREY JOYCE

WR-269-RWJF
May 2005
Prepared for the Robert Wood Johnson Foundation
This working paper contains technical material supporting a forthcoming Health Affairs article, “Costs and Effects of Participation in Collaboratives to Improve Chronic Illness Care” by Emmett Keeler and Geoffrey Joyce. It includes material that underlies the results in the article and was part of the journal submission. The technical material here was reviewed by the journal, but not by the formal technical review process at the RAND Corporation. It includes brief summaries of design decisions discussed at length in Cretin et al., 2004i, and data on participation and on item scaling from the various disease results papers summarized at http://www.rand.org/health/ICICE/findings.html. This working paper is intended to be read in conjunction with the article. We also report here some additional analyses completed later in the study to test for possible biases in the results.

Design

Our design goal was to compare the changes in measures at intervention sites and matched comparison sites, from before participation in the collaborative to after. We were able to find matched sites for all congestive heart failure (CHF) and diabetes sites. However, among the small asthma organizations, we could find only four comparison sites. We decided to allow an unbalanced design for that disorder and use the data from the seven unmatched intervention sites to supplement the four matched intervention sites.

Randomization is rarely used in studies of systemic change for good reason. Patients are not randomized because systemic changes are implemented for all the patients in the organization’s initial intervention site, and it would have been hard for these organizations to pick out a random half of patients to not include in the disease registry, not allow to go to group visits, not allow to use a new hotline, etc. Even at the organization level, randomization is rarely done. We were charged with evaluating a practical way to help organizations improve care. All participation was voluntary. Organizations that volunteered to participate, and in many cases pay a substantial fee to do so, wanted to improve care, but not to be research subjects. If we had tried to make them randomize their patients, or to have only half a chance of participating, we believe none would have participated.

So, unlike the many before–and-after case studies of one reforming organization, we had a before-and-after study with controls and many organizations. While the Chronic Care Model (CCM) is evidence based, in the sense that its components have been the subject of randomized controlled trials, prior studies of collaboratives and of the CCM as a whole were based on uncontrolled, self-reported results. The reason for our evaluation was to provide an external evaluation using a controlled design. Shojania and Grimshaw in their rather critical evaluation of the field were supportive of this design. ii

Organization Participation

The first collaborative we evaluated was run by the Institute for Healthcare Improvement (IHI) and covered CHF and diabetes. To increase the sample size, we added a large diabetes site from a regional collaborative run by the Washington State Department of Health and Qualis Health.
The next collaborative was also run by IHI and covered asthma and depression, with predominantly Bureau of Primary Health Care sites sponsored by the Health Resources and Services Administration. The last collaborative was another diabetes regional collaborative in Washington State. All collaboratives were co-led by Improving Chronic Illness Care.

At the outset, there were 107 organizations participating in the collaboratives, but 7 organizations dropped out: 1 CHF, 3 diabetes, 1 asthma, and 2 depression. Of the remaining organizations, 26 were excluded from the evaluation by our initial criteria: 24 diabetes organizations from the second Washington State Collaborative, which were too small to evaluate (we only considered internal comparison groups in that collaborative), and the 2 school-based asthma clinics, which were too different from the other asthma organizations to evaluate. This left 74 eligible organizations, of which we enrolled the following (the first number in the range is for patient data, and the second number in the range also includes organizations with staff surveys only):

- 6 to 10 of 14 CHF organizations
- 13 to 14 of 15 diabetes organizations
- 11 to 12 of 25 asthma organizations
- only 4 to 6 of 20 depression organizations (perhaps because of the sensitivity of this condition).

Both the asthma and depression organizations provided care to many undocumented aliens and were reluctant to risk scaring them with an unknown telephone caller. For the three diseases discussed in the article, the totals were 30 organizations with patient data and 36 organizations with staff surveys out of a total of 54 organizations. In most organizational analyses, we included interviews from the 6 depression organizations that completed staff surveys, for a total of 42. We made an effort along with the collaborative faculty’s efforts, to get collaborative organizations to enroll in the evaluation, but we could not force them to do so.

The sponsor covered collaborative tuition (i.e., the $12,000 fee to IHI for participation for a few “scholarship” organizations in the first collaborative) but did not support time spent at the collaborative or implementation of quality improvement at any organization. The scholarship organizations did not do as well according to the faculty assessments, and the sponsor stopped giving the scholarships. The evaluation was overlaid on an essentially voluntary activity. We agreed to cover any purely research costs, such as going through the institutional review board process and getting consents, and we paid small amounts for these to many of the organizations that participated.

**Selection of Comparison Groups**

In the collaboratives, organizations were encouraged to pick a site and an initial “pilot” group of 100–300 patients at that site with the disease of interest with whom to develop their system changes. These might be the patients with diabetes from one clinic or one group of doctors. We asked the organizations to identify another comparable clinic site or group of doctors in the organization that would not be immediately affected by the collaborative. In all but six cases, comparison groups were internal—the patients from another clinic site or group of doctors in the
organization who were not immediately affected by the collaborative. We believed that such comparability, including shocks to the organization as a whole and ease of getting permission to study both intervention and comparison patients from only one organization, outweighed the possible problem of control groups “looking” better than they should because of quality improvement ideas leaking from the intervention site to the comparison site.

**Patient Recruitment and Sample Size**

In Exhibit 1 in the Health Affairs article, we report the number of patients who filled out the first survey and were used in the patient survey analysis—after dropping patients (1) who consented but when contacted said, or whose charts said, they did not have the disease; (2) who said they normally did not receive care in either the intervention or control sites; or (3) who because of sickness or death could not do the survey. (Statistics on how we get from the sampling frame to the final sample are in the disease-specific papers). Although a few people who did not receive telephone surveys were included in the charts analyses, somewhat fewer people in all were used in the charts analysis: Consent procedures were not always identical, and more importantly, several of the process items related to any health service use within a one year period. Therefore, we selected only patients whose charts covered the 25 month period from 11 months before the start of the collaborative to 14 months after. The 11 months before the collaborative began was the “before period”; the first three months after the start were excluded to give the sites a chance to begin making changes; months 4 to 14 were the “after period.”

We cannot say exactly how many staff participated in staff surveys—681 filled out our first wave of paper surveys (anonymously), but in addition, we had structured telephone interviews on costs and on thoughts from a year later, including greatest successes and barriers and whether the changes lasted. Because the initial surveys were anonymous and about 75 percent of staff filled them out, there may be a few more who participated in our later structured phone interviews. In fact, the membership of the quality improvement (QI) teams changed during and after the intervention. For simplicity, we report that we interviewed 681 staff.

Because so few organizations in the depression collaborative were willing to give us patient data, we had no way to evaluate their processes and outcomes for depression patients independently and so have dropped them from the article, except for a footnote.

**Power**

The sample size for patients was large (approximately 2,000 intervention patients and 1,800 control patients), but the limited number of organizations (42 at most) prevents any complicated analysis of factors associated with success. Almost all of our analyses of processes and outcomes were within disease, but we still found many significant results, so power at the patient level was not a significant limitation. Power is discussed further in the evaluation design paper (Cretin et al., 2004), where we show that for CHF and asthma, we are powered to detect a 10 percent difference in our scales, and for diabetes, an 8 percent difference, pooling results across sites, but not across diseases.
Time Horizon of Evaluation

The collaborative lasted for one year. The after period for the charts lasted another three months. The sponsor wanted a quick assessment—we tried to assess longer-term results by follow-up interviews with QI team leaders a year after the end of the collaborative as reported in the text above and in the article. In theory, the impact of many of the systemic changes—such as new software to track and manage patients—would last. Such chart outcomes as cholesterol and blood pressure control predict future heart disease.

Assessing the Importance of Changes

We assessed which types of changes were likely to have more impact using a method developed for this study. This concept was called the “depth” of the change in Pearson et al., 2005

http://www.rand.org/health/ICICE/findings.html, where we provide a description of these methods and data on depth of changes, in addition to their number. In reality, the impact of “made a registry of patients,” “provided a hotline to a care manager,” or any of the other 90 changes in Pearson’s taxonomy of CCM system change depends on how it was done, what was happening before, and many other factors. The importance of the changes is reflected in the changes in patient process and outcomes that we measured.

Analytic Methods: Patient Analyses

Comparisons based on the charts, primarily of technical process, but also of some clinical outcomes such as heart disease risk factors for patients with diabetes, were differences in differences—i.e., differences before and after with a control group.

However, patient surveys (which cover current attitudes and care for the preceding six months) could not begin until the organizations agreed to be in the evaluation, completed institutional review board review, obtained patient consent, and provided a sampling frame of patients with contact information. The median delay from the first learning session to sending out the surveys lagged from the ideal 4–6 months to 10 months in IHI 1 (diabetes and CHF), and to 15 months in IHI 2 (asthma and depression). As a result, comparisons based on the surveys (patient knowledge, self-management, and utilization) were just intervention vs. comparison at approximately the end of the collaborative, with adjustment for observed factors.

In some cases, both charts and surveys asked about the same domains and complemented each other’s results. For example in CHF, the charts of patients with heart failure who were cared for at intervention sites had higher documented rates of counseling regarding medications, diet, exercise, smoking, weight control, and self-management activities compared with rates at control sites at the end of the collaborative process, but equal rates at the beginning. However, these latter findings were based on chart review alone, and intervention sites could have merely improved their documentation in these areas. Patient surveyed at the end reported more education about their disease and self-management, better communication with providers, and more shared decision-making. These data show that the observed chart changes are not just due
to better documentation. In addition, the similarity in baseline chart data shows that the reported differences between intervention patients and controls at the end of the collaborative probably represent changes due to the collaborative and not a continuation of better self-management in the intervention sites.

For processes and outcomes, we pooled across sites but not across diseases. Methods varied across the disease-specific papers, but we usually used hierarchical models with patients nested in sites. When we did not use hierarchical models, we adjusted the significance of the results for clustering by site. We had baseline data for variables derived from the charts, but not from the surveys.

**How Many Process and Outcomes Were Tested to Get the Results Shown in Exhibit 2?**

For asthma and CHF, we have papers on process and of patient reported self-management scales and outcomes. For Diabetes, we have written up process, clinical outcomes, but not our analysis of patient reported measures. In all three diseases, process and clinical outcomes were measured by 10-18 clinical indicators, which were arranged into 2-4 subscales and an overall process scale. Patient survey items were also grouped into scales of education, knowledge, communication, self-management behavior and health outcomes. In the disease specific papers, we also report individual items. We did not control significance for multiple comparisons because many scales and items were significantly improved.

**Methods for Attributing Non-Collaborative Costs**

We did not do cost analyses for the control sites. We asked the intervention organizations for the additional costs due to planning for or implementing the quality improvement. Often this was straightforward, such as the salary of a new care manager, or purchase and programming of new software, but in many cases respondents had to use their judgment for which parts of care were additional.

**Analysis of Cost Offsets**

To estimate cost offsets, because we were not able to get claims data from the organizations, we had to rely on patient interviews on their 6 month prior utilization. Because we did not have access to pre-intervention utilization data, we needed to adjust for as many factors as possible. Because visits, and hospitalization are count variables, we use negative binomial regression, controlling for an intervention dummy, disease severity, age dummies, education dummies, black, Hispanic, male, month of survey relative to the start of collaborative, adjusting significance for clustering by site. We used the coefficient estimates from the negative binomial
models to obtain the predicted number of visits or hospitalizations or hospitalizations per person for pilot and control patients.

The cost offsets are cost-savings to society, and not necessarily to the participating organization. How to organize health care and its financing to promote the higher quality that leads to such cost-savings is a complicated issue. There would be an immediate payoff for only one of the participating CHF organizations, but in principle, generating these cost-savings should have been advantageous to the others when they talk to insurance companies or employers about premiums and other contract conditions.

Methods for Organizational Analyses

For organizational analyses of the association between organizational attitudes and what was done, we used all 42 organizations (including 6 depression organizations) where staff was surveyed. All respondents (even the chronic care providers in the comparison groups) assessed their organization’s commitment to quality improvement using a modified Baldrige Quality Improvement Assessment (Shortell, et al. 1995), and intervention staff were asked their perceptions of the chronic care collaborative, motivation, and workgroup climate. In addition, a key contact completed an organizational characteristics instrument. In an follow-up survey, team members were also asked to rate their team effectiveness.

Most of the survey analyses did not use comparison groups, but looked at the associations of different measures within the 42 intervention groups comparing them to other survey measures and to the number and depth of changes that were made.

Possible Biases

Selection biases could be operating at three levels.

Across Organizational Levels

For generalization, only organizations that wanted to improve care signed up for the collaborative and in many cases paid substantial tuition, so we can not generalize to organizations that don’t want to improve. For a few measures, we could compare our results to trends in national data, and usually our comparison sites beat the national trends, showing an organization wide desire or readiness to improve. Because of matched comparison sites, this desire should not affect our results, but we don’t have information on whether collaboratives will help organizations that are not motivated to improve.

For internal validity, organizations in the Collaborative that expected to do well might be more likely to agree to participate in the evaluation. We tested for this as follows. The Collaborative faculty gave ratings of final success to all but 3 organizations that completed the collaborative on a scale from 1-5. 3 organizations (2 depression, 1 diabetes) had no final faculty rating of success.
39 of the remaining 71 organizations agreed to be in the evaluation, and we can compare faculty assessment for controls and participants. For all diseases but depression, 33 of 51 were in the evaluation. Overall and by disease, participating organizations did slightly better but the difference is neither practically nor significantly different. (an average of 0.11 better, p for difference = .35). If we leave out depression, the 33 participating organizations had an average of 4.1 the 18 non-participating organizations had 3.9, (p = 0.19). Also for each disease, the differences by participation were never significant. For CHF, the difference of .325 had p = .19; for asthma, the difference of .11 had a p of .64; for diabetes, the non-participating organization did better than the average of the participating organizations, for depression participating organizations were .02 better with p = .91. The faculty ratings for the depression organizations averaged 4.0, similar to the three diseases we do report on.

Within the Organizational Level

is also problematic for our design—clinics or staff that volunteered to go to the collaborative might be the most motivated to improve in the organization. Indeed, in many cases, staff at the intervention sites did volunteer. (It is part of QI philosophy to use enthusiastic people to build initial success that might inspire others to follow.) If so, their patients might have improved more than those from comparison sites without the collaborative. The comparison sites were supposed to be the ones that the organization planned to spread to later. Surprisingly, in our staff surveys of intervention and comparison site chronic care providers, we saw insignificantly less favorable assessments of quality improvement as measured by the Baldrige survey in the intervention site staff (for each of the 4 subscales we studied). That they made more changes or got more resources is not a bias—that is part of the intervention.

At the Patient Level

If the consenting intervention patients are healthier or more able to learn than the consenting comparison patients, then they will change more. There is no reason why factors that influence patients to consent to be in the study should differ between intervention and control site, particularly because only 6 of the 23 comparison sites were external with separate consent provisions. Perhaps as a result, differences in consent rates between intervention and control patients are small. We control for observable differences in age, disease history etc. in the analysis.

Control Site Improvement

The control sites, which almost always come from the same organization, may have improved more than they would have if the intervention site had not been in a collaborative to improve care. This could come about if staff or perhaps information technology leaks from the intervention site to the control site, or if the organization decides to spread the intervention
before we have completed our measurements. If so, control site comparisons will underestimate the improvement due to the collaborative.

In addition to collecting patient data from the control sites, we have a very limited number of paper surveys, and in addition we conducted a one-hour retrospective phone interview with 21 of the 23 control sites one year later. We asked the control staff person who was supposed to be most alert to quality improvement activities about overlap, and whether they were also doing quality improvement for that disease to help us in interpreting the patient results.

For analyses of the impact of overlap, we developed a scale that included geographic separation, overlap of physicians and of other staff. This scale was used in sensitivity analyses and was mildly related to process and outcomes of control staff patients.

The controls did not have a formal collaborative QI team for the disease, so did not make a monthly report of changes that they made. So we can not make a direct comparison of changes with the intervention site because of the methods differences. We do not know of data on how many of these systemic changes are made by normal organizations over the course of a year.

In general, 74 percent of the control sites reported implementing at least one intervention-related change during the collaborative period—62 percent of these changes were diffused from the intervention site, 43 percent from other sources (i.e., their own QI efforts, 7/9 of which were “minor” defined as specific QI activity related to individual CCM-related areas, such as guidelines, reminder systems, rather than comprehensive). The average depth rating (coded using Pearson’s 0–2 scale) was slightly higher for changes from the pilot site vs. other sources (1.22 vs. 1.09), but not enough to be statistically significant, with only n=21 sites.

For diabetes, we can look at national trends in some of our items using HEDIS or Veterans Affairs (VA) data. Especially in the first collaborative, the control sites improved much more than national trends, as our organizations caught up to the elite organizations represented in HEDIS and the VA.

In sum, there are many potential biases due to the design, but as far as we could check, none of them were serious in the event. In our analyses, we did not adjust for comparison site improvements being more than normal, so our results may be conservative for that reason.

---

