6. Discussion

Analysis of Published and Ongoing Research

Performance-based payment and reputational incentives. The available literature about QBP is sparse and there is little evidence base from which to answer the key questions listed in Chapter 2. For those studies that are available, the results are mixed. The incentive strategies used and dependent variables measured are too different among the studies to permit formal meta-analysis. Key variables frequently go unreported, making it more difficult to reach firm conclusions about the potential for and limitations of QBP. Furthermore, since several important variables are not included in any study, the potential for these factors to influence the observed results of these studies is unknown. This means users of the available studies of QBP must be cautious and rely on their judgment in drawing lessons from this literature.

With those caveats, it does appear that in some circumstances, providers respond appropriately to financial incentives. For instance, Hickson et al. show that a financial incentive as small as $2 per visit is enough to increase pediatrics residents' willingness to do well-child care and provide continuity of care.[43] Similarly, Hibbard et al. show that reputational incentives increase the quality improvement activities of hospitals, especially those that are performing poorly.[61]

The optimal approaches to QBP—and the determinants of when one approach is more effective than another—remain uncertain, given the literature. However, some factors identified in the conceptual model do seem to matter. In particular, the observation that significant responses were more likely for quality indicators that reflected clinician performance, rather than patient compliance (e.g., for tobacco screening vs. tobacco cessation) suggest that enabling or inhibiting factors are important. By extension, this implies that the difficulty and cost of achieving the performance goals (both of which rise when patient barriers increase) may also be important determinants of the response to an incentive.

In addition, both studies in which the performance threshold for receiving bonuses was uncertain (because it depended on the performance of other medical groups) were negative. This suggests that uncertainty about revenue potential may be a factor, but strong conclusions cannot be reached based on two studies, even though they are randomized and controlled.

Other potentially important factors have not been studied. These include potentially predisposing factors such as the presence and impact of other community initiatives or incentive programs (which could create an "incentive cacophony"), provider characteristics, and enabling (or inhibiting) factors at the organizational level.

The absence of studies of organizational factors may be particularly important, since responses to some incentives may be determined at the organizational level. For instance, many observers are advocating the use of clinical teams and information systems, both of which would be difficult and expensive for a single provider but might be more feasible for a group. Furthermore, an increasing number of providers are practicing in group settings,[4] so a rising proportion of the priority-setting for, and systems investments by, providers comes from the organizational level**.** The optimal approach to QBP may include a mix of incentives directed at both organizational and individual provider levels, but this has never been studied.

The focus on preventive measures in the available literature likely reflects that these have traditionally been disproportionately represented in accreditation and other data collection

processes (such as HEDIS®) and that they can be measured from administrative data, which is part of the reason they are in HEDIS® and also makes research easier to perform. However, this focus also means that most of the available literature addresses quality problems in the underuse category, rather than overuse or misuse, even though problems of this type are quite common.

Ongoing research into QBP. The research projects that are currently in progress will provide some important information. They should provide an estimate, at least at a point in time, of the extent of use of QBP, and describing several specific projects and the determinants of participation among providers. In addition, several evaluations using contemporaneous (though not randomized) control groups or natural experiments are planned, and these should provide some information about the impact of various QBP strategies. However, the lack of randomized controlled trials makes it extremely likely that there will be continuing questions about each of the QBP strategies tested, especially about whether uncontrolled factors that differ among the intervention and control groups explain the observed results. Moreover, simple trials are not designed to test the effects of a QBP intervention with sufficient sample size to assess whether performance differs across the various predisposing and enabling factors that might affect variations in performance. In essence, they would be analogous to testing whether chemotherapy "works" against cancer, without specifying the nature of the drugs, their regimens, or even the type and stage of the cancers.

Evaluating Outcomes Reports

Our simulations suggest that outcomes reports can yield useful evaluations of hospital and other provider performance. We reach different conclusions from prior investigators for three reasons. First, we assume that, while mislabeling may occur in a single period, it is unlikely to have significant impact on a hospital unless it is repeated over multiple years, which we show would be a very rare event. Second, we introduce the notion of several categories of providers. We believe that it is less important to mislabel a provider from its own category to the adjacent one than it is to miss by multiple categories, and we find that these major mistakes are rare, even with relatively small sample sizes. Finally, by using recent data reflecting the much larger than previously expected differences in outcomes (which have been validated by studies of processes), we have modeled hospital populations with larger differences in underlying mortality rates. These results are consistent with the notion that chance can have an impact on providers' reputations in the short term, but that it should not be a major barrier to outcome reporting if one assumes long term relationships between providers, their patients, local purchasers, and other stakeholders.

In addition, our results show that, despite the statistical "noise" created by random variation, evaluation and labeling systems can be developed that can discriminate poor quality hospitals from good or superior hospitals. Such evaluations, by their nature, will have better grading accuracy when the distributions of the underlying hospitals to be graded (that is, the groups of the hypothetical world) have little overlap. Overlap is reduced when scores are based on outcomes in which the difference between good and poor (or superior and good) performance is large or when the number of patients per hospital is large (to minimize variation due to chance). In cases in which the outcomes in question have overlapping distributions in the hypothetical world, the evaluation system can be improved by using multi-category evaluations (i.e. more than just the labels "good" and "poor") and summary grades over time. Each of these approaches has pros and cons.

Evaluation using multiple categories has the advantage that one can be more assured of accuracy of the grades that differ most from the mean. In our examples, it would be unlikely for a hospital with a 3-year *star* score of 3 (tentatively rated as *poor* 3 consecutive years) to actually be a superior quality hospital. However, multi-category grading does increase the chance of minor mislabeling. It would be a fairly common event for a hospital to receive, for example, *5 stars* in a given 3-year period, even when its long term performance was actually at the *6-star* level. In addition, how the multi-level category scores would be perceived by and interpreted by the users of hospital performance reports is an issue that requires careful thought. In our hypothetical hospital domains, there would not be a reliable difference between hospitals receiving a score of 5 or 6. Yet some stakeholders may tend to order hospitals with these middle scores, despite the lack of reliable differences among them. This is not an uncommon situation in other scoring systems—e.g., *Consumer Reports* frequently indicates that certain products are of approximately the same quality and are listed alphabetically.

Multi-category scores are perhaps most useful in that they can identify a subset of hospitals that are almost definitely truly *superior* or truly *poor* quality. With the former group, one can search for process differences that could form the basis of benchmarking or providing lessons for process improvement at other hospitals. Several processes contributing to improved outcomes were identified by analyzing the reasons for outcome variations among hospitals in New York.[70] Conversely, hospitals with definitely *poor* performance can be studied to search for process-level explanations for their sub-par outcomes. Thus, measuring outcome data may help us learn which processes to change and monitor. Furthermore, hospitals should not, and are not likely to, wait until they receive three consecutive *poor* quality assessments before doing something. While a single *poor* score may just be chance, any reasonable quality improvement team would start examining charts and processes after a second *poor* quality score, if only to be able to report back to the CEO on what they found before the next quality reports come out.

Future Research

Study design issues. From the literature review, it should be clear that pursuing research without a conceptual/theoretical model leads to incomplete reporting of key variables, and research designs that produce results that are not very useful for policy recommendations. Thus, the first requirement for subsequent research should be a clear delineation of how it fits into an overall scheme for testing conceptual models of QBP. For instance, a common (and valid) target for quality improvement is cancer screening. In the short term, cancer screening can be expected to increase utilization (by both the initial testing and the evaluation of positive tests). In most capitation agreements, these additional services would be covered by the capitation fee. Therefore, theory suggests that any evaluation of a QBP program to increase cancer screening in capitated environments should explicitly consider the magnitude of the costs of screening for, diagnosing, and treating more cancer in comparison to the incentive offered. This is not to suggest that the simple economic incentive within capitation to avoid screening costs leads to bad behavior by providers focused on their capitation balance. However, if the organization's quality improvement committee is trying to decide whether to focus limited resources on responding to an incentive to increase screening for cancer or an equivalent incentive to increase physician counseling regarding smoking cessation, the latter might be chosen, because it is less

burdensome in a variety of ways. Only by considering all these costs and barriers to responding to an incentive can its true impact be understood.

There are also important issues of control groups and analytic plans. Though there clearly need to be more randomized controlled trials of QBP, these are difficult and often expensive to undertake. However, as the number of purchasers and health plans adopting QBP increases, there will be more opportunity to use contemporaneous control groups that, though not randomly selected, could be useful, especially if attempts are made to match them to intervention groups in terms of characteristics identified in our conceptual model. As these study designs are more subject to bias than randomized trials, we believe extensive use of qualitative analytic methods will be valuable in augmenting the quantitative analysis of an incentive's impact with participants' and observers' judgments about barriers to and determinants of responses to the incentives.

Topics of investigation. Theory also should play a greater role in the selection of topics to be studied. Since most of the existing research focuses on incentives to individual providers, but the conceptual model suggests that organizations could have a profound influence on performance, a topic needing further investigation is the relative importance of individual versus organizational incentives. In addition, the model suggests the need to address special situations, such as when market characteristics (e.g., local monopolies) are the dominant feature of purchaser-provider relations. This does not imply that all studies must begin with theory—we recognize that in many instances researchers will have to work with the interventions that are being put into place by purchasers. Theory, however, may help inform the selection of intervention goals, of the timing of site involvement, and of the selection of "control" or comparison groups. The theoretical framework we have outlined may also help design better interventions simply by causing people to think more carefully about the incentives, enabling factors, and potential barriers.

Finally, we found only one trial that compared two different QBP approaches; all other studies had a "placebo" control group. A major goal should be to address this weakness with studies that compare performance-based payment to reputational strategies and compare different strategies within the payment and reputational subcategories to each other. These evaluations should include temporal components as well. For instance, it may be that there is some attenuation of response to reputational strategies over time if they are not subsequently backed up with payment incentives.[5]

Planning research programs. While individual research projects should reflect theory, funders may also wish to consider using theory to drive their approach to developing a portfolio of research. In particular, we suggest two general approaches, which we refer to as *sequential hypothesis testing* of incentive strategies and *parallel hypothesis testing* of enabling and predisposing factors. By sequential hypothesis testing we mean that a research program could proceed in a logical fashion from tests of incentives that have a higher probability of being successful (that is, of stimulating performance improvements) toward those that, *a priori*, would be expected to be less likely to be effective.

For instance, consider the QBP strategies of additional fee-for-service payment versus paying bonuses to providers from a fixed pool based on relative performance. There are features about the bonus pool approach that purchasers might find attractive, such as: 1) the total payout can be set in advance, 2) purchasers can raise or lower this figure periodically and precisely as provider performance and market situations change (e.g., after initial investments to improve performance are paid off, providers may need less of an incentive to continue or to make smaller incremental

gains), and 3) other stakeholders can see exactly how large a commitment purchasers are making to quality. On the other hand, most of the current health care environment is fee-for-service and providers may be more willing to accept, or at least more likely to believe they understand the implications of, a fee-for-service approach. They may also be resistant to the program if they feel that even if they improve, their bonus would be jeopardized if someone else improves more—especially if they could argue that the data or baseline states were not comparable. Therefore, it may be reasonable, as a supporter of research, to consider fee-for-service projects that seem feasible initially, with the understanding that even if these work, it does not guarantee that other methods with which providers are less familiar will have similar impact. Thus, a strategy of simply finding whatever works in getting providers used to being "measured" and receiving explicit rewards for improved performance may be more important than finding the "best" QBP method, at least initially.

If findings accrue suggesting incentive programs that had a higher *a priori* chance of success are indeed effective, funders could begin to consider projects that at least initially seem less feasible or less likely to succeed. If the results of these subsequent trials are negative, they do not negate the prior results, but help place bounds on which approaches are effective. On the other hand, if the subsequent trials are positive, they suggest a wide variety of incentive strategies may be useful. Alternatively, if the approaches thought to be most effective do not work, then either they were "sub-clinical dose", or the underlying strategy should be re-thought. Similarly, the absolute magnitude of the incentive may be an issue, in which case it is useful to start with high pre-test probability of success (that is, with fairly large incentives) and move progressively lower to understand what magnitude of incentive is needed to change behavior. In this manner, the field could move sequentially along a spectrum of hypotheses within each conceptual domain of incentive characteristics, delimiting the range along which QBP strategies can succeed.

Understanding the key aspects of alternative incentive approaches will be important, but will take some time. Therefore, we also recommend simultaneous assessment of the impact of the other elements of the conceptual model, predisposing and enabling factors that mediate the response to incentives. For instance, it is very likely that predisposing factors such as the general financial environment and enabling factors at the organizational level will influence performance, regardless of the use (or not) of QBP strategies. To enhance our understanding of both the potential of QBP and the settings in which it is effective, funders might consider supporting parallel programs addressing these other elements of the conceptual model. That is, getting organizations to install improved information systems, or revising the economic incentives against the coordination of care and preventive services, may by themselves be sufficient to lead to improved performance, without any specific QBP incentives.

It may also be useful to consider the results of this parallel research into predisposing and enabling factors when evaluating subsequent QBP proposals. For instance, if research showed that organizations with disease registries had consistently superior performance, funders might consider whether subsequent QBP trials should be limited to organizations that have registries for the conditions for which performance is measured. It is also important to recognize that certain features may be crucial for some interventions and not others. Registries may be critical to assure that appropriate care is given to patients with diabetes or hypertension, because insufficient contact with the medical care system may be especially problematic for these patients; registries are unlikely to be needed to assure that beta blockers and aspirin are appropriately recommended upon discharge after a myocardial infarction. This combination of

sequential hypothesis testing of incentive strategies with parallel hypothesis testing of other elements of the conceptual model is likely to advance the field much more rapidly than has occurred to date.

The basic tools of performance measurement. Another barrier to QBP is that the science of performance measurement is still underdeveloped.[5, 71] The available set of metrics is not broadly representative of all care, while purchasers must pay for care across the entire clinical spectrum. Furthermore, there has been little research addressing non-clinical outcomes such as absenteeism that may be very important to employer purchasers.[71] Experience in other industries has shown that developing performance measures for complex phenomena is difficult and that inappropriate measures can have significant negative consequences.[72] This suggests that research into QBP should be accompanied by further development of the basic tools of performance measurement.

Conclusion

The environment in which purchasers and providers interact is rapidly changing. There is clearly growing interest in QBP and some evidence that both payment and reputational incentives can work, but, to date, there is little unequivocal data on which to base QBP strategy selection. Fortunately, our modeling suggests that, with appropriate caution, outcomes measures can be included among the performance indicators used for QBP. Furthermore, the notion of using incentives to encourage high quality (as well as actually measuring quality) is much more acceptable than it was a few years ago, and this has increased the number of opportunities to study QBP. Researchers have responded with a broad portfolio of ongoing research that promises to both outline current trends in the use of QBP and offer some preliminary evaluations of several different incentive approaches. Policymakers should expect additional research, especially if designed and selected for funding based on conceptual considerations such as those we outline, to rapidly advance our understanding of how to use performance measurement and incentives to improve the quality of health care Americans receive.