January 31, 2013

Quality Control for the IOM Study of Geographic Variation in Health Expenditures

Draft Final Report

Prepared for

Institute of Medicine
Robin P. Graham, PhD, MPH
Senior Program Officer
Institute of Medicine
National Academy of Sciences
500 Fifth Street, NW
Washington, DC 20001

Prepared by

Thomas J. Hoerger, PhD
Benjamin Yarnoff, PhD
Simon Neuwahl, MA
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194

RTI Project Number 0213602

RTI Project Number
0213602

Quality Control for the IOM Study of Geographic Variation in Health Expenditures

Draft Final Report

January 31, 2013

Prepared for

Institute of Medicine
Robin P. Graham, PhD, MPH
Senior Program Officer
Institute of Medicine
National Academy of Sciences
500 Fifth Street, NW
Washington, DC 20001

Prepared by

Thomas J. Hoerger, PhD
Benjamin Yarnoff, PhD
Simon Neuwahl, MA
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194

Contents

Table

1. PURPOSE OF THE STUDY

The purpose of this report is to provide quality control to the Institute of Medicine (IOM) study of geographic variation in health expenditures. As part of this project, RTI International performed audits of the qualitative and quantitative data synthesis performed by IOM subcontractor Precision Health Economics (PHE) and the microsimulation impact modeling performed by IOM subcontractor RAND Corporation (RAND). We reviewed the subcontractors' work in five general areas, which are described in further detail in Section 3 (Methods):

1. Data sources and data processing
2. Sampling
3. Measurement
4. Statistical analysis
5. Results and conclusions

2. MAJOR FINDINGS FROM THE QUALITY AUDIT

2.1 PHE

The PHE report synthesized the results from three separate studies commissioned by IOM that focused on separate payer populations (Medicare and Medicaid, performed by Acumen, LLC; commercial payers in the MarketScan database, performed by Harvard University; and commercial payers in the OptumInsight database, performed by the Lewin Group). PHE also constructed and analyzed a measure of total population health care spending, including commercial, Medicare, Medicaid, and the uninsured.

We found that PHE performed all of the tasks in its statement of work and generally applied the assumptions and specifications provided by the IOM Committee. We were able to replicate all of the PHE analyses following directions contained in the report or other documentation and program files provided by PHE. In general, our replications produced the same results as those contained in the PHE Final Report. In a few cases, PHE corrected typographical errors based on a preliminary list of discrepancies we provided. In other cases, our results remained slightly different from PHE's, but these differences were small (e.g., a coefficient of variation [CV] differed by 0.001) and did not affect PHE's conclusions. Overall, we characterize the degree of reproducibility as high.

We believe that the study provides a valuable synthesis of the findings from the individual subcontractors. The synthesis of results between payers fully answered the three key study questions related to the synthesis. Although we have minor concerns about the calculation

of total health care spending, we believe that the analysis adequately answers two of the study's three key questions on total spending; it is not clear whether the study answers the third question: Is total spending or Medicare-only spending a better predictor of Medicare quality outcomes?

Overall, we believe that the report passes the quality control process. That said, we have four remaining concerns about the Final Report:

- One of the most notable findings in the report is the high CV for Medicaid spending. This finding may be related in part to variation in the Medicaid fee-for-service (FFS) share between states. The Medicaid FFS share varies widely between states, and spending is generally much higher for FFS beneficiaries than for beneficiaries in Medicaid health maintenance organizations (HMOs). As a result, Medicaid spending in a state with a high Medicaid FFS share may not be comparable to spending in a state with a low Medicaid FFS share. This comment should not be construed as a criticism of the analyses performed by PHE. PHE was asked to synthesize the results from the commissioned Medicaid study, and the authors have correctly analyzed the Medicaid spending data they were given. However, we believe it is worth talking about the limitations of the Medicaid data when discussing the Medicaid results.

- In constructing the total spending variable, PHE set low and high outliers in the Medicaid HMO to non-HMO spending ratio equal to the mean value for the ratio. Although setting the low outliers to the mean ratio has little effect on the constructed total spending variable, we believe that adjustment of the high outliers will have a larger effect and is inappropriate.

- We are concerned that the authors are overinterpreting the results on geographic variation within hospital referral regions (HRRs) and subsequently overemphasizing some of their more speculative arguments in the Discussion section. The report concludes with the recommendation that "future work in the area of variation in health care might focus less on geography per se, and more on the contributions of individual provider and hospital behavior, and incentives, to the variation that is observed in spending and utilization" (p. 30). We do not believe that this discussion adds much to the report, and—because it is in the concluding paragraph of the report—it may actually detract attention from the rest of the study's findings. In some ways, it almost sounds like we should turn our backs on the study's findings on geographic variation in order to look at individual providers. That is probably not what the authors intended, and, based on the study findings, we do not think such a conclusion is warranted.

- The Discussion section should briefly discuss study limitations.

On January 13, 2013, IOM's Ashna Kibria asked us to look at the numbers in Table 13 of the PHE report. Specifically, she noted that for analyses of variation in HSA level spending, some members of the IOM Committee had been concerned that "some of those numbers, for example, the OptumInsight or Harvard numbers, seem a bit off. The percent attributable is much higher than what was found in the original subcontractors' reports." We examined PHE's method and calculations as well as the methods and numbers reported by Harvard and Lewin in their reports. We were able to replicate the PHE results in Table 13, using the

datasets they received from the subcontractors. The methods used and results reported differed between PHE, Harvard, and Lewin, so it was difficult to reconcile the numbers between reports. When we looked at the OptumInsight hospital service area (HSA)-level spending data that PHE analyzed, we noticed that some HSAs were either extremely high (>$1,000 pmpm) or extremely low outliers (<$100 pmpm). The extreme outliers usually occurred in HSAs with relatively few observations. In an HSA with few OptumInsight patients, one patient with high spending or a string of especially healthy patients could have a large effect on the HSA mean. In an unweighted regression, these HSAs will account for much of the overall variation in HSA spending. To confirm this, we reran the PHE specification, first omitting the 129 HSAs with less than 100 observations and then omitting the 243 HSAs with less than 200 observations. The share of HSA spending variation explained by HRRs increased from 5% with no HSA omissions to 25% with the 129 HSA omissions and to 31% with the 243 HSA omissions.

Our concerns and specific comments on the PHE report are discussed in more detail in Section 4.

2.2 RAND

RAND modeled the potential impact of three policies (bundled payment, pay for performance [P4P], and accountable care organizations [ACOs]) on variation in Medicare spending across HRRs. The authors estimated 2008 Medicare spending for each HRR under the baseline case and various scenarios for each policy. They found that neither P4P nor ACOs would have a substantial impact on geographic variation in Medicare spending, but bundled payment would decrease geographic variation in spending for the care included in the bundles.

RAND accomplished three separate and complicated modeling exercises under a very tight period of performance. At the highest level, we believe that the report's ranking of the comparative effects of the three policies on geographic variation is reasonable and correct (i.e., bundled payment reduces geographic variation, whereas P4P and ACOs have little or no impact). The authors used separate models for each analysis. We considered whether it would have been feasible—given the period of performance and available evidence—to use a common modeling framework for all three analyses. Although it would not have been feasible to develop such a framework from economic first principles or determinants of geographic variation (because these determinants are not known), we believe it would have been possible and helpful to at least develop an explicit mathematical formulation for spending that would provide greater insight into how the different policies affect the components of spending.

The report generally does a good job of modeling multiple kinds of payment and documenting the complicated process of computing payment. The analyses generally make

limited assumptions about behavioral responses to the incentives included in the new payment policies, under the rationale that there is limited evidence of behavioral responses. However, the three policies are relatively new and have not yet been adopted on a wide scale; moreover, lack of a measured response does not mean that no behavioral response exists. The ACO simulation appears to be the most exploratory, because ACOs are the newest policy and have the least evidence.

The Conclusions section briefly summarizes the results and provides intuition for why the results differ between policies. The section does not try to overinterpret the results, properly leaving much of the interpretation to the IOM Committee. Still, a little more comparison of the results across policies might be helpful in explaining why the bundling simulation produces a larger effect on geographic variation than P4P or ACOs. The section should also briefly discuss key limitations of the study. In the last paragraph, the section correctly emphasizes that even though the P4P and ACO policies do not reduce geographic variation, that does not mean that the policies are ineffective. Neither of these policies has an explicit goal of reducing geographic variation, and the improvements in quality and/or efficiency associated with each policy are important in their own right. These concerns and specific comments on the RAND report are discussed in more detail in Section 5.

3. METHODS

We applied the same general approach to assess the quality of the reports from PHE and RAND, although some details of our approach varied based on the objectives and content of the two reports. For both projects, we obtained the subcontractors' statement of work, correspondence reflecting IOM's guidance and specifications, and relevant materials from the IOM Web portal for the overall geographic variation project. Our assessment is based on the following reports:

- Precision Health Economics (PHE). (December 13, 2012). *Geographic variation in health care spending and promotion of high-value care*. Final report. Prepared for the Institute of Medicine.
- RAND Corporation. (October 2012). *IOM Committee on Geographic Variation in Health Care Spending and Promotion of High-Value Care: A modeling of policy recommendations*. Draft report. Prepared for the Institute of Medicine.

We met with each subcontractor to obtain additional data and discuss assumptions for their analyses. Below, we describe details of our approaches.

3.1 Data Sources and Data Processing

For both reports, we reviewed databases used by the subcontractors; described the rationale for using those databases and their overall generalizability; reported on the reliability and validity of the databases, quality checks (e.g., missing data, coding errors),

and data cleaning procedures (where indicated); and detailed the subcontractors' decision making and methods for addressing error.

The subcontractors (particularly PHE) worked with geographic variation data inputs generated by other IOM modeling teams. We reviewed the input data sets to assess whether the subcontractors correctly incorporated the data and any modeling assumptions determined by the IOM Committee. We assessed whether missing data exist and were accounted for in the analysis. We reviewed the subcontractors' methodology in deciding how to incorporate the data and identified any steps where the decision making was unclear or questionable.

3.2 Sampling

RTI reviewed and reported on the selection plan, inclusion and exclusion criteria, and steps applied in deriving the final sample from the initial population. IOM provided us with the key spreadsheets from the IOM modeling teams that were provided to the data synthesis and microsimulation modeling teams. These spreadsheets included information on expenditures by HRRs. We reviewed the subcontractors' final reports to identify which variables were included or excluded and the rationale for these decisions.

3.3 Measurement

RTI reviewed variables specified by the subcontractors. In particular, this review investigated the reliability and validity of variable definitions. We identified whether the subcontractors included the variables most appropriate for identifying geographic variation in health expenditures and assessed whether these variables were defined in standard ways.

3.4 Statistical Analysis

For PHE's data synthesis analyses, we assessed the validity of methods applied to the descriptive summary of "across modeler" methods, qualitative synthesis of "across modeler" regression modeling results, and HRR spending analyses. PHE synthesized information from three other IOM modeler groups and examined whether variation in expenditures is similar across Medicare beneficiaries, Medicaid beneficiaries, and privately insured individuals. To assess the validity of the synthesis efforts, we evaluated whether we could replicate results by following the steps described in the PHE Final Report.

We assessed whether the data synthesis subcontractor's methods were sufficient to identify similarities and differences between results in the across modeler comparisons. For example, we examined whether model results were adequately compared and contrasted based on the differences in their study populations. For the qualitative synthesis, we evaluated whether the rationale for qualitative conclusions was clearly explained.

For RAND's microsimulation modeling exercise, we examined justification for and effects of variable omission on analyses, described missing data for model variables, and explained how missing data were addressed analytically. We also assessed whether RAND performed sensitivity analyses and tested statistical assumptions appropriately.

We reviewed the basic structure and key assumptions of the microsimulation models used by RAND to simulate alternative reimbursement policies. Because RAND developed separate analyses for each policy, we assessed whether a single, comprehensive modeling framework could have been used to analyze the different policies. We examined whether the analyses included behavioral responses to changes in reimbursement policy, and we evaluated whether other potential behavioral responses would be likely to affect the simulated outcomes. We identified major parameters expected from theory and assessed whether these were included appropriately in the models. We assessed whether the subcontractor followed standard practices for microsimulation modeling.

We assessed whether the microsimulation results had face validity (i.e., did the results make sense?) and consistency (e.g., if more variable expenditure patterns were input into the model, did the simulations produce more variation in results?). In the case where the predicted results were unexpected, we examined how the analysis led to the unexpected results. Unexpected results are not necessarily wrong, but it should be possible to explain how the results are generated by the model. Because sensitivity modeling and uncertainty analysis are key components of simulation modeling, we identified which variables were varied, the range of variation for each variable, whether additional key input variables should be varied, and whether probabilistic sensitivity analyses should be performed. If probabilistic sensitivity analyses were performed, we assessed whether appropriate distributions were applied.

3.5 Results and Conclusions

To evaluate the subcontractors' results and conclusions, we evaluated intermediate summary output and final results. We examined whether the subcontractors' findings were internally consistent and assessed whether the subcontractors clearly stated their criteria for making conclusions and that the results supported their conclusions. We reviewed the limitations stated by the subcontractors, identified any additional limitations, and—to the extent possible—discussed whether these limitations were likely to change the subcontractors' key conclusions. If the subcontractors made policy recommendations, we considered whether the recommendations are fully supported by their findings or whether additional study may be needed.

4. RESULTS OF THE QUALITY AUDIT OF THE PHE REPORT

The PHE report synthesized the results from three separate studies commissioned by IOM that focused on separate payer populations (Medicare and Medicaid, performed by Acumen, LLC; commercial payers in the MarketScan database, performed by Harvard; and commercial payers in the OptumInsight database, performed by the Lewin Group). The synthesis attempted to answer the following questions:

- Across the studies, how do spending, utilization, and quality vary within and between areas?
- How much variation is explained by observed predictors? Are the explanations consistent across different payers and populations?
- At what level(s) of geography is the variation occurring?

PHE also constructed and analyzed a measure of total population health care spending, including commercial, Medicare, Medicaid, and the uninsured. This analysis attempted to answer the following questions:

- How does total spending vary across regions?
- What predictors explain this variation?
- Is total spending or Medicare-only spending a better predictor of Medicare quality outcomes?

Briefly, PHE concluded that significant variation exists in spending across regions, and health status and other measured factors cannot explain all of this variability. Spending across different payers is not perfectly correlated; there are fairly modest correlations in regional utilization patterns across payer. Similarly, regional variation in quality is also not well-correlated across payers. These facts suggest that providers might be responding differently to patients with different sources of coverage. The evidence also suggests that variation exists across relatively small regions. Based on the results, PHE (2012) concludes that "future work in the area of variation in health care might focus less on geography per se, and more on the contributions of individual provider and hospital behavior, and incentives, to the variation that is observed in spending and utilization" (p. 31).

4.1 Overview of Quality Audit Methods

To begin the quality audit of the PHE Report, RTI spoke with research staff at IOM and described initial reactions to PHE's work. IOM staff referred RTI to their online portal where key results and methods documents from the original subcontractors were kept. Additionally, IOM invited RTI to request any relevant Committee correspondence (with the subcontractors and PHE) that might clarify methods. Using PHE's scope of work, RTI first

reviewed the current draft of PHE's report to identify any missing tasks or unclear sections. Some were identified, and RTI followed up on these with IOM and PHE prior to making conclusions about the quality of PHE's work.

For many of these ambiguities or missing tasks, IOM provided clarifying documents, such as PHE's presentations to the Committee, technical reports from subcontractors, and correspondence with the Committee or IOM staff. This step cleared up some omissions from the Final Report (e.g., analysis of significant variables across payers [Task 2.2] was included in a presentation, and the use of InterStudy data in place of the American Community Survey [Task 3.1.A] was approved by the Committee).

Next, RTI reviewed PHE's descriptive summary of subcontractor work and found it to be more succinct and cohesive than the "master methods document" provided on IOM's portal. Although RTI did not engage in a full review of individual subcontractor reports, some reports were later reviewed incidental to the replication of PHE's synthesis results (e.g., Harvard's technical report). No discrepancies were noted between PHE and subcontractor descriptions during this process. Additionally, RTI engaged Norma Gavin, a primary contributor to the IMPAQ Quality Control review of the subcontractor results. She provided us with further detail and insights into the subcontractor methods and how they differed. Again, RTI found PHE's summary of subcontractor methods to be accurate, without being overly detailed. It was clear from Dr. Gavin that a closer discussion of details and differences across subcontractors was available in the IMPAQ report.

The process of replicating PHE's synthesis results was relatively simple. Because this process only required an HRR-level area effects file from each subcontractor and a few Excel functions, RTI did not examine any program logs for PHE's synthesis results. Identical results produced using two distinct methods would satisfy the quality control process. However, several discrepancies were found during replication. These were documented and sent to PHE with IOM copied.

Most of these discrepancies were very small. Several were simply typographical errors that PHE corrected, while others resulted from PHE using older versions of subcontractor results (although these were likely the most recent results at the time PHE produced its report). The only discrepancy of mild significance was a typographical error in Table 2 that listed MarketScan's coefficient of variation as 0.15 instead of 0.12. All typographical errors and discrepancies due to older versions were corrected by PHE and documented by e-mail, with IOM copied.

Subsequent to replication of the synthesis work by PHE, RTI determined that program logs and further documentation would be needed to replicate the more complicated policy synthesis and meta-analysis (Task 3 in PHE's statement of work). RTI held a call with PHE

and IOM staff prior to the interim report deadline to clarify key aspects of their analysis and submit a request for documentation and program logs showing the details of their work on Task 3. Because of the quantity of data sources, data management exercises, and transformations involved in this task, RTI took an "understand and verify" approach rather than replicating results from the raw files (some of which were protected by data use agreements).

Because of their complexity and vulnerability to calculation or programming errors, RTI was particularly interested in PHE's work on estimating Medicaid managed care spending, uninsured spending, and total spending. RTI received a timely response from PHE containing a large set of Stata program files, log files, and some of the raw files. An overview of how these various files related to each other allowed RTI to "understand and verify" that PHE stuck to the methods described on pages 24 through 26 of the Final Report (Total Spending—Methods) and that no programming errors were made.

Although RTI did not replicate the total spending results from scratch, subsequent to the "understand and verify" process, RTI requested final datasets from the total spending analysis in order to replicate the calculation of area means, standard deviations, and CVs, similar to the replication process completed for the subcontractor synthesis. We used these datasets to replicate the total spending analyses in the PHE report.

Because of the importance of Medicaid managed care data in estimating total spending and the sensitivity of results to these data, RTI sought to confirm that the appropriate data had been extracted from the Medicaid Statistical Information System (MSIS). RTI became familiar with the MSIS by reading CMS' documentation and by re-extracting the same data described by PHE. Although these data had some limitations (e.g., FFS "eligible persons" had to be compared with HMO "enrolled persons" when calculating per capita spending), we agree that PHE executed this Medicaid managed care analysis appropriately, given the data. Similarly, RTI sought to carefully understand and verify PHE's uninsured analysis and use of Medical Expenditure Panel Survey (MEPS) data. To complete this task, RTI examined the MEPS variables used in log files. Because of RTI's familiarity with MEPS data, this process was very simple. Although no deviations from their stated process were found, RTI felt that PHE had not engaged the full utility of these data (see notes on MEPS health status variables in total spending section under "Specific Comments on the PHE Final Report"). We also tested the assumption that national averages for inpatient/outpatient care can be applied to any region (for application of the input price adjustment to uninsured data) and found that state averages do not vary widely; thus, results are not sensitive to this assumption.

On January 13, 2012, IOM's Ashna Kibria asked us to examine the results in Table 13, because some IOM Committee members were concerned that these results were not

consistent with the results reported by the subcontractors. We reviewed the subcontractors' Final Reports and attempted to explain any differences in methods and results between the subcontractors and PHE. Results of this review are presented in Section 4.4.6.

4.2 Overall Findings

Overall, we believe that the report passes the quality control process. Specific findings include the following:

- We were able to replicate almost all of the results in the report.
- PHE performed all of the tasks specified in the statement of work. Some appear in their presentations to the Committee, but not in the Final Report. We presume this was discussed with and approved by IOM and the Committee.
- PHE appears to have followed Committee guidance in creating variables and performing analyses.
- The study provides a valuable synthesis of the findings from the individual subcontractors.

That said, we have four remaining concerns and a number of specific comments about the Final Report, which we discuss in detail below. The length of our discussion should not be compared to our assessment (above) of the report's accomplishments, because it takes less space to explain that we agree with a study's data and methods.

4.3 Major Concerns

4.3.1 Medicaid Variation

One of the most notable findings in the report is the high CV for Medicaid spending. The wide variation may be due in part to differences in Medicaid managed care penetration rates (specifically for HMOs) across states (with corresponding effects across HRRs). The Population and Databases section correctly notes that the Medicaid managed care population represents more than half of Medicaid beneficiaries. Two other facts are relevant, but not mentioned:

- Medicaid managed care enrollees typically spend less than Medicaid FFS enrollees. According to the Kaiser Family Foundation article cited in the report, managed care covers about two-thirds of Medicaid enrollees and accounts for about one-third of Medicaid spending. The article attributes the differences in spending to differences in the populations covered by managed care and FFS, with managed care predominantly covering children and FFS often covering high-cost beneficiaries.
- The HMO share of Medicaid enrollees varies widely between states, ranging from 0% to more than 80%.

The first fact implies that the mean reported per member per month spending for Medicaid (FFS) will substantially overstate average Medicaid spending for all beneficiaries. It is worth stating this as an explicit limitation of the study, or at least mentioning it when discussing the mean outcomes of spending measures across populations in Table 3. This would state the limitation more strongly and clearly than the current limitation that appears just before Methods: "As a result of this exclusion, the study results may not generalize to the total Medicaid population" (PHE, 2012, p. 11).

The second fact complicates interpretation of the reported Medicaid CV, which is much higher than the CVs of other payers. To the extent that Medicaid managed care enrollment is associated with low-cost Medicaid enrollees, the difference in penetration rates across states means that we will not be making apples-to-apples comparisons. The reported Medicaid spending rate for a state (or HRR) with a 0% Medicaid HMO penetration rate will cover the full cost range of Medicaid enrollees, whereas the reported Medicaid spending rate for a state (or HRR) with a 50% Medicaid HMO penetration rate is likely to be higher because it is based on a disproportionate share of high-cost patients. These differences may occur even if average Medicaid spending is the same in both states.

We emphasize that these comments should not be construed as a criticism of the analyses performed by PHE. PHE was asked to synthesize the results from the subcontractors, and the authors have correctly analyzed the Medicaid results they were given. That said, we believe it is worth talking about the following limitations of the studies when discussing the Medicaid results:

- Mention the two related facts when discussing Medicaid in the Population and Databases section.
- When discussing Table 3, note that the spending rate for Medicaid FFS is probably much higher than the rate for the overall Medicaid population, so we should not necessarily conclude from the table that Medicaid enrollees are most expensive.
- Similarly, when discussing the CVs in Table 5, note the possible complications arising from differences in managed care penetration causing differences in FFS severity across states. In principle, correcting for age and health status might partly address this problem, although the percentage reduction in CV from controlling for these factors is not as big as it is for Medicare. Market factors, including managed care penetration rates, do not have much of an effect on the Medicaid CV, but that is probably because overall managed care rates are used instead of the Medicaid managed care rates that are more relevant from the standpoint of the patient severity issue.
- Consider discussing this issue in conjunction with the huge drop in the Medicaid CV associated with Cluster 10. This drop merits a longer discussion than it currently receives because the drop is so large, and it is not matched by a similar drop for Medicare. It would be interesting to know whether the institutionalization variable

has a big effect on the CV, because that might be correlated with the share of patients who are in managed care.

- Construction of the Total Spending variable offers an opportunity for looking at variation in Medicaid spending for all patients. Construction of Total Spending requires estimating Medicaid HMO spending per beneficiary and Medicaid FFS spending per beneficiary. These estimates can be weighted by their corresponding patient shares to provide an estimate of the average Medicaid spending for all (both HMO and FFS) beneficiaries. While replicating the Total Spending analysis estimates, we performed a simple analysis of the variation in Medicaid spending at the state level. We found that the CV fell from 0.431 based on Medicaid FFS beneficiaries only to 0.319 based on all (HMO and FFS) Medicaid beneficiaries. (Our simple analysis did not correct for outliers on the HMO to non-HMO ratio. For more on outliers, see Section 4.3.2.)

4.3.2 Total Spending

Outliers in the Medicaid HMO to non-HMO spending rate. Data on relative spending between Medicaid HMO and non-HMO beneficiaries are not available at the HRR level. Therefore, PHE relies on state-level estimates from the MSIS Data Cubes. This is an appropriate choice. However, we have concerns about the decision to set low and high outliers equal to the mean HMO to non-HMO spending ratio. When the ratio is less than the 10th percentile or higher than the 90th percentile, PHE sets the ratio equal to the mean ratio (48.63%). We believe that this approach is acceptable for the 4 states with ratios below the 10th percentile (12.4%) because (a) these states have relatively few HMO beneficiaries, (b) the ratios imply that HMO spending may be implausibly lower than (less than one-eighth of) non-HMO spending, and (c) the small HMO share means that the mean ratio will have relatively little effect on average spending for all Medicaid beneficiaries. In contrast, the 4 states (Arizona, Hawaii, New Mexico, and Tennessee) with ratios above the 90th percentile (95.6%) all have large HMO shares (45% to 82%), and their HMO spending is plausible (ranging from roughly equal to 1.5 times larger) relative to non-HMO spending. Consequently, in these states, applying the outlier adjustment would have a substantial effect on the constructed combined Medicaid spending value. Therefore, we recommend not substituting the mean value for these high outliers.

Finally, for the 14 states with $0 HMO spending in the MSIS Data Cubes, the HMO to non-HMO spending ratio is set equal to the mean value for the ratio. This is an acceptable way to deal with the zero values because the HMO weight for most HRRs in these states will be close to or equal to zero, and thus the assumed mean ratio will have little effect on average Medicaid spending for all beneficiaries. Currently, however, Appendix 2 Table B is misleading because it shows the distribution of Medicaid managed care to Medicaid FFS ratios after adjusting for both the outliers and the zero values. It would be clearer to show the original distribution of non-zero values for the 37 states and then show the distribution after adjusting for outliers and zero values. Treatment of the zero values could be explained in the table notes.

4.3.3 Interpretation of Variation within Areas

We are concerned that the authors are overinterpreting the results on geographic variation within HRRs and subsequently overemphasizing some of their more speculative arguments in the Discussion section. This emphasis begins with the findings and discussion on Variation within Areas on pages 22 and 23, continues with the last two paragraphs of discussion under "Does Total Spending Predict Medicare Quality Better Than Medicare Spending?," and culminates with the closing two paragraphs of the Discussion (and the whole report), where the authors recommend that "future work in the area of variation in health care might focus less on geography per se, and more on the contributions of individual provider and hospital behavior, and incentives, to the variation that is observed in spending and utilization" (p. 31).

The authors seem to base their argument on the finding that there is variation in spending within the HSAs in an HRR, so that any policy that focuses on reducing variation across HRRs may still leave variation within the HRR. Although this finding is accurate, we believe the authors overemphasize the 50% of HSA variation that is not explained by HRRs and underemphasize the 50% of variation that is explained by HRRs. Reducing that 50% of variation attributable to HRRs might be a goal for policy makers (e.g., Congress) who are more concerned about how providers in an area fare and less concerned about individual providers. Regional-based policies might also be more feasible to implement than policies aimed at individual providers.

Beyond the policy perspective, we do not believe that this discussion adds much to the report, and—because it comes in the final paragraph of the report—it may actually detract attention from the rest of the study's findings. In some ways, it almost sounds like we should turn our backs on the study's findings on geographic variation in order to look at individual providers. That is probably not what the authors intended, and, based on the study findings, we do not think this implicit conclusion is warranted. Rather than putting their recommendation in either/or terms, the authors might say, "In addition to looking at variation in spending at the geographic level, we should also study variation in spending within geographic areas. By doing so, we may gain further insights about the factors that determine variation in spending at the geographic level."

4.3.4 Study Limitations

The Discussion section should briefly discuss study limitations.

4.4 Specific Comments on the PHE Final Report, 12/13/2012 Version

4.4.1 Executive Summary

- Second paragraph, first bullet: Standard deviations of quality outcomes provide less meaning than CVs would.

- Third paragraph, first bullet, second sentence: Change "that" to "than."

4.4.2 Background on Subcontractor Data

- First paragraph: Report MarketScan covered lives.

- Page 10, description of Medicaid: See earlier discussion of Medicaid FFS population and differences in Medicaid FFS fractions between states that may complicate comparisons between states.

4.4.3 Findings of Analysis

- We were able to replicate results in Tables 5 through 12. We also replicated Table 13, based on the datasets provided by the subcontractors to PHE. But see the related bullets below and in Section 4.4.6's longer discussion of Table 13.

- There is no Table 14 in the current version of the Final Report. In a previous version of the report, Table 14 contained cohort analyses that were dropped from the current version. PHE will renumber the subsequent tables when it finalizes the report.

- Table 5: The large decrease in the CV for Medicaid associated with Cluster 10 probably deserves greater discussion.

- Variation within Areas, page 22, first paragraph: "This regression of HSA level outcomes with HRR level random effects isolates the share of variation in mean spending and utilization outcomes that is occurring at the HRR level. By subtracting that share of variation from 1, PHE can pinpoint how much variation in these outcome measures is attributable to the geographically smaller HSAs. In effect, ***it*** provides an upper bound on how much variation a regionally-targeted policy could hope to reduce." This section becomes a bit confusing. By the time we reach the last sentence, the reader has a hard time knowing whether "it" refers to (a) the share of variation in mean spending and utilization outcomes that is occurring at the HRR level, (b) 1 minus that share of variation, or (c) how much variation in these outcome measures is attributable to the geographically smaller HSAs (i.e., how much HSA variation is left after controlling for HRRs). We think (b) is the same as (c), but it is difficult to tell. We think (a) represents the upper bound on how much variation a regionally targeted policy could hope to reduce, but the section could make that clearer.

- Table 13. Share of Variation in HRRs Attributable to HSAs: Is the table title correct? We believe that a more accurate title would be "Share of Variation in HSAs That Is Not Attributable to HRRs" or "Share of Variation in HSAs That Is Not Explained by HRRs" because the underlying analysis regresses HSA spending and utilization while including random effects for HRRs. This is closely related to the confusing section in the first paragraph on page 22. If the table title is corrected, the second paragraph under Variation within Areas on page 22 will have to be rewritten. In Section 4.4.6,

we note that PHE reports the HSA variation in spending results differently from the subcontractors. PHE may enhance comparisons between results by adopting the same reporting format.

- Next paragraph, extending from page 22 to page 23: As noted under our major concern about interpretation, this paragraph appears to overinterpret the result that variation still occurs within fairly defined geographic areas (presumably, they mean HSAs) even after controlling for HRRs. The overinterpretation occurs when they move from the finding that some variation remains to argue that geography might be incidental to the source of the variation. However, some remaining variation does not imply that none of the variation occurs at the HRR level. If the authors' argument was correct, we would have expected that including HRR random effects would have little effect on HSA variation. In fact, it appears that the variation is cut in half. Thus, a better interpretation of Table 13 is that HRRs explain a significant portion of HSA variation, but about half of the variation remains. This still leaves room "for deeper analysis into the presence and causes of variation across individual physicians and hospitals," but it does not mean that future work on geographic variation should be neglected.

4.4.4 Total Spending

- Page 24, last complete paragraph: "A limitation of the MarketScan data is that area spending is censored for 53 of 306 HRRs." The reason for censoring should probably be explained in the report. In their programs, log files, and HRR-level total spending dataset, 58 HRRs are missing total spending data. The censoring appears to be related to restrictions in the data use agreement between Harvard and PHE. If 53 additional areas are censored for the total spending variable, does this mean that the CVs for MarketScan in the previous sections are only shown for 253 (or 248) areas? If so, it might be worth mentioning this in the previous sections and noting that we are comparing the MarketScan CVs for 253 areas to the Medicare and Medicaid CVs for 306 areas. It would also be reassuring to know that the Medicare and Medicaid CVs for the 253 areas included for all payers are similar to the Medicare and Medicaid CVs for the 306 areas.

- Page 25, section on control for outliers in the ratio of HMO to non-HMO cost per Medicaid enrollee: The control method may be masking true differences in HMO and non-HMO costs. This could be especially problematic if the ratio is correlated with the Medicaid managed care share (see Major Concern 2—Total Spending in Section 4.3.2).

- We were able to replicate Tables 15 through 20. After reviewing RTI's comments on the replication of Tables 19 and 20, PHE decided to change its specification for the quality estimates. PHE's quality regressions had measured Medicare spending using the component that gets added into total spending. That differs from the Medicare spending from Acumen, because, for example, the PHE variable multiplies the Acumen variable by the HRR share of the population in Medicare. On reflection, PHE decided it is more natural to use the Medicare FFS spending from Acumen and plans to update the report accordingly. We concur with PHE's decision.

- Table 15, predictors included in clusters for total spending regressions: One remaining source of variation in total spending is variation in payer shares between HRRs. For example, an HRR with a high proportion of Medicare patients will likely have higher total spending per person than an HRR with relatively few Medicare

patients. Because PHE has already calculated spending by payer, it would be possible to create an estimate of payer share-standardized total spending by applying national average payer shares to spending by payer in each HRR. PHE was not originally asked to run this analysis, but it may provide useful information to the Committee.

- Page 27: PHE mentions that "because the health status predictors were specific to a particular study or population, those predictors were additionally weighted by that population's share of the total HRR population." However, they do not mention which health status predictor—if any—was used for the uninsured population. Census region and Metropolitan Statistical Area (MSA)/non-MSA specific health status measures might have been calculated from MEPS. These include the Physical Health Component Score and the Mental Health Component Score (only one of these would likely be chosen). However, it is not clear whether enough uninsured observations are included in MEPS. Ideally, PHE would say whether a health status variable was included for the uninsured and, if so, which variable was used.

- Table 19: The text states that "the IQI composite was better predicted by total spending." However, it appears in the table that the R-squared values for all three quality variables are always higher when Medicare spending is included instead of total spending.

- Table 20: The coefficients listed under Medicare spending for specification 4 are identical to the coefficients for Health Status (Medicare) in Appendix 3. PHE confirmed that the variable name in Appendix 3 is wrong and that three health status variables should appear in that table. The table will need to be revised to reflect the new Medicare spending variable included in the updated regression.

- Page 31: Discussion of the "striking findings" in Table 20. This discussion may be a little too long, relative to the discussion for other findings in the report. The reasoning here seems relatively speculative, and the amount of space it receives, relative to other comments, may cause readers to give the discussion more emphasis than it deserves. Also, the following sentence appears to be wrong: "However, one also has to explain why Medicare spending is associated with lower quality for two of the three measures, and higher quality for the third." The negative coefficients signal higher quality, according to the last paragraph on page 30.

4.4.5 Discussion

- The first paragraph correctly notes that the analysis "does not shed light on whether variation is valuable or harmful" (p. 31). We quibble, however, with the second example where specialization means that the variation is valuable. For variation to be valuable in this case, two factors must hold: (a) there must be some reason why high-intensity care is easy to attain in the high-intensity sections of the country but more difficult to obtain in the low-intensity sections of the country, and (b) outcomes in the high-intensity section of the country must be better enough to justify the added costs associated with high-intensity care.

4.4.6 Table 13 Results

On January 13, Ashna Kibria asked us to look at the numbers in Table 13, noting that "There has been some concern from the Committee that some of those numbers, for example, the OptumInsight or Harvard numbers, seem a bit off. The percent attributable is

much higher than what was found in the original [subcontractors] reports." We examined PHE's methods and calculations, as well as the methods and numbers reported by Harvard and Lewin in their Final Reports (we did not find similar analyses in the Acumen Final Report). We did not have the datasets that Harvard and Lewin analyzed, so we could not replicate their results, nor could we examine whether they would have obtained the same results if they had followed PHE's methods (or vice-versa). As a result, it is difficult to pinpoint exactly why the results in the PHE report appear to differ from those in the other reports. However, we can offer some insights into the question (see especially the sub-bullet under the Lewin results):

- We were able to replicate the PHE results in Table 13, using the dataset PHE received from the subcontractors.

- The methods used and results reported differed between PHE, Harvard, and Lewin (see Table 4-1).

- The methods used by PHE and Harvard seemed to be most similar. Both basically estimated the share of variation in HSA-level spending that was attributable to HRRs. However, PHE reported the share of variation in HSA spending that was not attributable to HRRs, whereas Harvard reported the share of variation in HSA spending that was attributable to HRRs. Simple comparison of the two numbers is misleading because the share of variation in HSA spending that is not attributable to HRRs equals 1 minus the share of variation in HSA spending that is attributable to HRRs.

- When we set both equal to the share of variation in HSA-level spending that was attributable to HRRs, the numbers based on the MarketScan data were 0.47 for PHE compared with 0.70 for Harvard.

- We cannot rule out that this difference was caused by (a) differences in methods or (b) differences in the version of the dataset (it is possible that PHE analyzed an earlier dataset than the final dataset used by Harvard).

- The Lewin results are not directly comparable to the PHE results. Lewin concluded that their analyses "revealed statistically significant variation in PMPM spending at the HSA level that was not captured in an HRR-level fixed effects regression. However, the magnitude of the variation at an HSA level within HRRs was relatively low for the large majority of HRRs." Although Lewin's conclusions are broadly consistent with PHE's general conclusion that about 50% of HSA variation is not explained by HRRs, it is harder to square Lewin's other finding of significant HRR effects with PHE's finding that 95% of the variation in OptumInsight spending at the HSA level is not explained by HRRs.

  – One partial explanation for PHE's OptumInsight results in Table 13 is that the Stata command (xtreg) that PHE uses to compute random and fixed effects does not allow for HSA weights based on the number of observations within each HSA. When we looked at the OptumInsight HSA-level spending data that PHE analyzed, we noticed that some HSAs were either extremely high (>$1,000 pmpm) or extremely low outliers (<$100 pmpm). The extreme outliers usually occurred in HSAs with relatively few observations. In an HSA with few

OptumInsight patients, one patient with high spending or a string of especially healthy patients could have a large effect on the HSA mean. In an unweighted regression, these HSAs will account for much of the overall variation in HSA spending. To confirm this, we reran the PHE specification, first omitting the 129 HSAs with less than 100 observations and then omitting the 243 HSAs with less than 200 observations. The share of HSA spending variation explained by HRRs increased from 5% with no HSA omissions to 25% with the 129 HSA omissions and to 31% with the 243 HSA omissions. (If we put this in the format of HSA variation not explained by HRRs that PHE reported in Table 13, the values would decrease from 95% to 75% and 69%, respectively.) With these omissions, which partially mimic a weighting procedure, PHE's estimates move in the direction of Lewin's conclusion that the magnitude of "within" variation in an HRR is small.

  - In the MarketScan data, omitting HSAs with less than 100 or 200 patients did not eliminate as many HSAs, and these omissions had relatively little effect on the Table 13 values. MarketScan has more covered lives than OptumInsight, so it is not surprising that fewer HSAs were eliminated in the MarketScan exercise. We note that Harvard included HSA weights in its fixed effects estimation. It appears that HSA weights cannot be applied in Stata's xtreg command with random or fixed effects (HRR weights would be possible, but not helpful for this problem), so Harvard may have used SAS for this analysis.

- As noted in our draft final report, we had concerns about (a) how PHE described the results they presented (i.e., confusion about whether they were presenting variation that was or was not attributable to HRRs), and (b) their title for Table 13 (we thought that "Share of Variation in HRRs Attributable to HSAs" should have been "Share of Variation in HSAs That Is Not Attributable to HRRs"). Alternatively, to enhance comparability with the subcontractors' reports, PHE may want to present its results in the same manner as the subcontractors. Thus, PHE could report "Share of Variation in HSA Spending that is Attributable to HRRs" ("Attributable to" could also be replaced by "Associated with" or "Explained by"). If the results are reported this way, the new values would be equal to 1 minus the current values.

- One of our main concerns with the PHE report is the authors' interpretation of the results. We are concerned that the authors overinterpreted the results on geographic variation within HRRs and subsequently overemphasized some of their more speculative arguments in the Discussion section. The authors seem to base their argument on the finding that there is variation in spending within the HSAs in an HRR, so that any policy that focuses on reducing variation across HRRs may still leave variation within the HRR. Although this finding is accurate, we believe the authors overemphasize the 50% of HSA variation that is not explained by HRRs and underemphasize the 50% of variation that is explained by HRRs. The overemphasis continues in the Conclusions section to the point where it almost feels like they are saying we should give up on studying geographic variation and instead focus on individual providers. This overemphasis could be why the Committee viewed the PHE results to be so different from the private payers.

Table 4-1. Differences between Studies

Item	PHE	Harvard	Lewin
Dependent variable	Input price-adjusted spending at HSA level (also unadjusted spending and utilization)	Input price-adjusted spending at HSA level	Individual spending
Estimation	Random effects (fixed effects give similar results)	Fixed effects	Fixed effects for HRRs and HSAs
Data	Average costs for 3-year period	Not clear from report	Data for 3 years
Weighting	No weighting	Weighted "by size of HSA"	No weighting, bigger HSAs have more individuals
Reported result	1 - rho = 1 minus the share of variation in HSA spending that is attributable to HRRs	Variation in HSA spending that is explained by variation in HRR spending	Significance of HSA fixed effects after controlling for HRR, CVs of HSA spending within each HRR
Reported 1 – rho			
Harvard	0.53	0.30	Not applicable
Lewin	0.95	Not applicable	Not reported
Reported variation in spending that is attributable to HRRs (i.e., rho)			
Harvard	0.47	0.70	Not applicable
Lewin	0.05	Not applicable	Not reported
Interpretation of results	"As Table 13 shows, about half of all variation in unadjusted spending in HRRs was actually occurring at the HSA level, across payers. That share rises for input price adjusted spending for commercial populations" (P. 21). "This finding suggests that variation occurs within fairly tightly defined geographic areas. Moreover, it is consistent with an even stronger hypothesis, that variation occurs at the level of individual providers—physicians and hospitals. If true, one would expect variation to exist across geographic regions, but geography itself might be incidental to the source of variation. Indeed, any grouping of physicians and hospitals would produce variation in this case. If true, it calls for deeper analysis into the presence and causes of variation across individual physicians and hospitals" (p. 21-22).	"Around 70% of the variation in HSA spending is explained by variation in HRR spending, when weighted by the size of the HSAs. Moreover, the standard deviation in HSA spending within HRRs is $236 (28% of total variation at the HSA level)." (P. 46).	Analysis "revealed statistically significant variation in PMPM spending at the HSA level that was not captured in an HRR-level fixed effects regression. However, the magnitude of the variation at an HSA level within HRRs was relatively low for the large majority of HRRs" (p. 70). "The results illustrate that while there was variation among HSAs within the HRRs, the level of dispersion within each HRR was generally quite low, with an average CoV of approximately 0.15" (p. 71). One of the main findings in the overall report is that "There is statistically significant variation in spending at the HSA level that is not captured by HRR-level fixed effects" (p. 3).

(continued)

Table 4-1. Differences between Studies (continued)

Item	PHE	Harvard	Lewin
RTI comments	We were able to replicate PHE's results in Table 13, using the dataset they received from the subcontractors. As noted in our draft final report, we had concerns about (a) how they described the results they presented, (b) their title for Table 13 (we thought that "Share of Variation in HRRs Attributable to HSAs" should have been "Share of Variation in HSAs That Is Not Attributable to HRRs," and (c) their interpretation of the results. We are concerned that the authors overinterpreted the results on geographic variation within HRRs and subsequently overemphasized some of their more speculative arguments in the Discussion section.		

5. RESULTS OF THE QUALITY AUDIT OF THE RAND REPORT

RAND modeled the effects of three policies that potentially could affect geographic variation in Medicare spending: P4P, bundled payment, and ACOs. RAND analyzed the impact of these policies on overall Medicare spending and variation in spending across HRRs. For each policy, RAND modeled several scenarios that differed based on the overall effect on total Medicare spending, the amount of spending that would be redistributed between providers, and assumptions about behavioral responses to the policies. Under P4P, many providers had large increases or reductions in reimbursement, but there was no systematic relationship between quality measures and spending per beneficiary. Therefore, geographic variation was not affected. For bundled payment, overall Medicare spending fell, and there were significant reductions in HRR-level variation in spending for the services included in the bundle and smaller reductions in variation for total spending (the bundled conditions only represented about 17% of total Medicare spending). For ACOs, there was little clustering of ACOs in high-cost areas. Therefore, the policy had almost no impact on geographic variation.

In our quality review, we focused on the following areas:

- the general modeling approach applied across policies,

- whether Committee specifications or assumptions were incorporated in the models,
- assumptions about policy effects on payment,
- assumptions about behavioral responses,
- checking results for each model, and
- checking whether study conclusions follow from results.

We recognize that RAND accomplished three separate and complicated modeling exercises under a very tight period of performance. At the highest level, we believe that the report's ranking of the comparative effects of the three policies on geographic variation is reasonable and correct. Below, we identify key assumptions and conclusions that we agree with as well as areas of concern; we necessarily provide more discussion on the areas with concerns.

5.1 General Modeling Approach

In assessing the RAND report, we first considered whether a single conceptual modeling framework would be (a) helpful for comparing and contrasting results across policies, and (b) feasible to construct. We first considered whether it would be possible to build a model from economic first principles (e.g., from provider incentives, cost functions, patient demand, etc.). Although such an approach is theoretically appealing, it would require much stronger evidence on the appropriate specifications for objective functions, costs, and demand than is currently available. Even if these specifications were available, we believe it would not have been feasible to construct such a model in the short time available for the study.

A second potential modeling framework would start with the major factors contributing to geographic variation and examine how a policy likely would affect each factor. This approach requires a clear understanding of the factors causing geographic variation. As we understand it, the IOM Committee has not reached clear consensus on these factors. Therefore, it would have been difficult for RAND to follow such an approach.

A third approach would start with a mathematical equation for geographic variation (or the CV) in spending and examine how a policy affects each component of the equation. For example, spending in an HRR for episode j could be determined by the number of episodes, the number of services per episode, and the payment per service (or, under some policies, episode). Total spending in the HRR would be calculated by summing across episodes, and variation across HRRs could then be calculated in the usual way. The effects of a policy on each component of the mathematical equation could be specified and the overall effect on variation estimated. The advantage of explicitly following this approach is that the differences between policies could be clearly identified and related to the results, providing

better intuition for why the policies have different effects on variation. We believe that this approach is feasible; indeed, RAND has implicitly followed a similar approach separately for each policy to estimate the effects on spending variation. Making the variation equation more explicit and clearly showing the effects of each policy on the equation's components could provide better intuition for the results and help ensure that the policies are assessed in a systematic way.

5.2 Pay for Performance (P4P)[1]

5.2.1 Payment Assumptions

The authors generally do a good job of modeling multiple kinds of payments and documenting the complicated process of computing payment based on P4P. The assumptions in Appendix A-1 represent an appropriate approach for implementing the P4P payments. However, the discussion of quality scores for hospital payments could be expanded to provide a better justification of the methodology. For hospitals, the authors did not compute full quality scores as in nursing home care and home health care, but instead used the Medicare value based purchasing scalar. They state that this circumvents the need for computing quality scores, but it is not clear why this is the case. A fuller explanation would provide more confidence in this methodology.

We have the following minor comments on payment assumptions:

- Using payment data from 2008 and quality data from later years raises validity questions but is unavoidable due to the lack of quality data from 2008.
- In Table 2, Scenarios 2 and 3 appear to be identical. We think that Scenario 2 should be checked as "Conservative program—2%" instead of "Robust program—15%."
- The formula on page 39 for nursing home improvement would be clearer if it included i and m subscripts as in the formula for home health on page 37.

5.2.2 Behavioral Assumptions

The authors note that there is sparse literature on behavioral responses to P4P incentives. Because of this, they only include one model that simulates a behavioral response, and this response is relatively limited. This approach is suitable for a conservative analysis, but it should be noted that the lack of a measured behavioral response in the literature does not mean that no behavioral response exists. The literature is relatively sparse overall because P4P has not been used widely yet. Thus, the lack of behavioral response in the literature could simply reflect the limited nature of the literature itself. So, although the conservative approach used in the analysis provides the best guess from the literature, it fails to address hypothetical questions about the behavioral assumptions under which P4P could reduce

[1] In this and the following modeling sections, we combine comments on the corresponding chapter and appendix for the model.

geographic variation. For example, the limited behavioral response that is modeled may be too small in the context of the robust (15%) incentive policy. Also, it could be interesting to examine the behavioral assumption that changes in quality are associated with initial spending levels. Under such assumptions, P4P may reduce geographic variation. This type of what-if analysis is interesting and appropriate in microsimulation studies.

A further issue is that even when the model allows for behavioral change, it does not model the effect of quality improvement on spending. For example, quality improvement may decrease readmissions, which would decrease Medicare inpatient spending. It may be that there is no literature on the effect of quality changes on future spending, which should be noted if so. Regardless, this is a notable omission and should be discussed.

5.2.3 Results

The Results section does an excellent job of demonstrating that P4P will not affect geographic variation. Figure 2-1 is a particularly nice way to highlight the fact that there is little correlation between current quality and current inpatient spending, and that the lack of this relationship leads to little change in geographic variation under P4P. We suspect that the same relationship holds for home health and nursing homes; it may be worth mentioning this. The description under Table 2.4 clearly explains why the impact of P4P is smaller for HRRs than for individual providers.

We have the following minor comments on the Results section:

- In Table 2.2, indicate what the percentiles represent.
- In the text describing Table 2.2, mention that the results are for Scenario 3 (the scenario with the most providers receiving increases).

5.2.4 Summary

The authors do a good job emphasizing that because there is no association between current spending and currently used quality measures, P4P does not affect geographic variation. If they have identified alternative quality measures, it might be interesting (although not necessary) to mention them specifically here. The summary also correctly notes that some newer P4P initiatives include efficiency or cost measures. Implied, but left unsaid, is the notion that such initiatives might lead to reductions in geographic variation. It might be helpful to make this explicit.

The summary might discuss two other factors. First, the P4P measures that are modeled only affect inpatient, home health, and nursing home components of Medicare spending. Therefore, they will not reduce geographic variation in the other components of spending. Second, the initiatives do not directly affect physician decision making. This is important

because variations in physician practice have been suggested as a possible explanation for geographic variation.

5.3 Bundled Payment

5.3.1 Payment Assumptions

The authors do a good job computing the median historical payment for all settings and use this as the overall national rate. However, they incorrectly compute the national rate for each setting, by computing the average proportion of total episode payments in each care setting and multiplying that by the median national payment. This is incorrect, because, mathematically, the median of the sum of the setting payments (i.e., the overall national rate) is not equal to the sum of the medians of the setting payments. If they had used mean payments rather than medians, this would be an appropriate approach, but with median payment it is incorrect. It is unclear why they do not simply compute the median payment for each setting using the actual claims data. It is also not clear whether this will affect the conclusions of the analysis.

5.3.2 Behavioral Assumptions

The authors assume that providers will not alter behavior in a way that (1) changes the amount Medicare pays, (2) changes the volume or mix of bundles provided, and (3) changes the mix and quantity of services used outside the bundle. This is a strong assumption that likely influences the results. As with the P4P behavioral assumptions, this assumption is made because of the lack of literature on the subject. However, the sparse literature on behavioral responses does not mean that no response exists. The lack of testing alternative assumptions leaves unanswered questions about the effect of bundled payments under alternative behavioral responses. This is especially problematic here, because without behavioral response, a bundled payment policy reduces geographic variation largely by assumption.

Behavioral assumptions are discussed in the first two full paragraphs on page 14. In the first paragraph on potential behavioral responses, it would probably be good to lead with the intended effect of bundled payments: to provide better incentives for choosing the efficient mix of services.

5.3.3 Results

The authors do a good job presenting results about the effect of bundled payment on geographic variation and on specific areas within each bundle for each condition. Looking carefully at the bundles and what drives variation is interesting and well done.

It may be worth noting that although the national rate in Scenario 1 is based on a 5% discount off the national median Medicare payment amount, this actually leads to a much

larger discount compared with the mean payment amount (in Table 3.6, this leads to a 15% reduction in spending for S1; the effect is even larger for S7). This suggests that the mean payment is greater than the median, and there will be a greater reduction in spending for high spending HRRs than the increase in spending (relative to a 5% reduction) for low spending HRRs. From a policy standpoint, a 15% reduction in spending per bundle might be unrealistic politically; the behavioral assumptions also may not hold for such a large reduction in spending.

We note that Table 3.6 in this version is identical to Table 3.5 in the earlier version for all numbers except the CV and 75th/25th ratios for S5. The authors should confirm that this change is intended.

We have the following minor comment on the Results section:

- The first sentence of the text below Table 3.6 refers to Table 3.5. This should be updated to refer to Table 3.6.

5.3.4 Summary

The discussion is a little misleading in the way that it states the conditions under which bundled payments would reduce geographic variation. The model assumptions largely guarantee that bundled payments will reduce geographic variation, and this is not clearly stated in the discussion. Bundled payment rates based on national averages (or medians) will necessarily reduce variation in payments. For example, with no behavioral responses, the reduction in geographic variation under Scenarios 1 and 2 depends only on the variation in the distribution of conditions treated per beneficiary. We think that it is important to make this clear in the discussion of results.

The authors present interesting evidence on the geographic differences in the components of bundles. The report would benefit from a fuller discussion of what these differences mean and what they can help explain. The authors may need to simply say that this is an important avenue for future research, but the results are so interesting that we feel some additional discussion is necessary.

5.4 Affordable Care Organizations

5.4.1 Payment Assumptions

The assumption that ACOs will reduce spending by 1% uniformly across regions is a strong one and has the potential to substantially influence the results. As with other policies, this assumption is made because of a lack of evidence. Again, it would be interesting to examine how sensitive the results are to this assumption. For example, differential reduction in spending by initial spending level is another pathway through which ACOs could reduce

geographic variation. It would be interesting to examine this hypothesis in a sensitivity analysis, because it may significantly alter results.

One of the major challenges in this modeling exercise is the difference in timing for Medicare spending (2008) and ACOs (mostly 2012). Therefore, there is little direct evidence on the effects of ACOs on actual spending. The authors can examine whether spending levels in the 2008 might affect subsequent ACO formation, but there is little evidence on how the formation of ACOs affects subsequent spending. The authors base their assumptions on spending effects on ACO program details.

We have the following minor comment on this section:

- In Table 4.1, Scenarios 4 and 5 appear identical. Add another component to distinguish between the two.

5.4.2 Behavioral Assumptions

The authors assume that ACO coverage is predicted only by the observable characteristics that are included in their regression model. This may be an important assumption if there are unobserved provider or beneficiary characteristics that predict participation in ACOs. For example, efficient providers or more healthy patients may select into ACOs. These effects may affect results, so it is important to state them as assumptions.

5.4.3 Results

The ACO results appear to be the most exploratory results in this report, because ACOs are the newest policy and consequently have the least evidence. The authors do a good job of presenting results so that the reader can see the location and effect of ACOs relative to spending.

5.4.4 Summary

The authors do a good job of discussing key results and implications of ACO expansion. The last sentence—"In any case, because of the relatively low participation in ACOs combined with the relatively small expected impact on Medicare spending, the ultimate effect on geographic variation is likely to remain small" (p. 27)—concisely summarizes the results of the simulation.

5.5 Conclusions

The Conclusions section (Chapter 5) briefly summarizes the results and provides intuition for why the results differ between policies. The section does not try to overinterpret the results, properly leaving much of the interpretation to the IOM Committee. Still, a little more comparison of the results across policies might be helpful:

- The P4P policy maintains the basic FFS reimbursement system but makes payment depend in part on quality measures. Therefore, the effects of the policy on geographic variation will depend on the extent to which the quality measures are associated with spending levels in the HRR. Because there appears to be little association, the overall effect on geographic variation is minimal. As modeled, total spending does not change, although some providers receive modestly higher payments and some receive modestly lower payments. Physician services are not directly covered by the policy, which could be significant if differences in physician practice patterns drive geographic variation.

- Bundling changes the unit of payment from individual services to episodes of care and bases payment in part or in whole on the national median payments for the episodes. These changes eliminate the effect of geographic variation in the number of services per episode on HRR spending, thereby lowering geographic variation in spending. As modeled, the policy produces large reductions in overall spending for bundled goods under most scenarios. Physician services are directly affected by the policy.

- ACOs give providers incentives to reduce spending, although the general FFS system is maintained. The effects of the policy on geographic variation will depend on the extent to which ACO penetration is associated with HRR spending levels. Because there appears to be little association between ACO penetration in 2012 and HRR spending in 2008, the overall effect on geographic variation is miniscule. The modeled overall reductions in total spending are modest. Physicians are covered by ACOs, although the effects of ACOs on physician decisions and practice patterns are not explicitly modeled.

The Conclusions section should briefly discuss key limitations of the study, particularly relating to behavioral assumptions. In most scenarios, no or limited behavioral responses are assumed; however, each of the policies is designed to change provider behavior.

In the last paragraph, the section correctly emphasizes that the P4P and ACO policies are not necessarily ineffective, simply because they do not reduce geographic variation. Neither of these policies has an explicit goal of reducing geographic variation, and the improvements in quality and/or efficiency associated with each policy are important in their own right.

We have the following minor comment on this section:

- The summary of results in Table 5.1 tends to minimize the impact of policies on total Medicare spending per beneficiary because each policy only affects a fraction of total spending. For example, the bundled payment policy has a substantial impact on spending for bundled episodes, but bundled services only affect about 17% of total Medicare spending. Similarly, the potential effect of ACOs on total spending is limited by the limited ACO penetration.

REFERENCES

Precision Health Economics (PHE). (December 13, 2012). *Geographic variation in health care spending and promotion of high-value care*. Final report. Prepared for the Institute of Medicine.

RAND Corporation. (October 2012). *IOM Committee on Geographic Variation in Health Care Spending and Promotion of High-Value Care: A modeling of policy recommendations*. Draft report. Prepared for the Institute of Medicine.

PRECISION
HEALTH ECONOMICS
11100 Santa Monica Boulevard, Suite 500 | Los Angeles, CA 90025
Phone: 310.982.6310 | Fax: 310.982.6311 | www.PrecisionHealthEconomics.com

Geographic Variation in Health Care Spending and Promotion of High-Value Care

Response to RTI Review of PHE Work

6/28/2013

PRECISION HEALTH ECONOMICS

This document responds to the RTI International's report assessing PHE's work for the IOM study "Geographic Variation in Health Care Spending and Promotion of High-Value Care." PHE appreciates RTI's thoughtful and constructive review of its work.

Through the course of interactions with RTI and the IOM, PHE addressed nearly all of the issues raised in RTI's report, and updated the February, 2013 version of the report that was posted on the IOM study's website. For example, Table 13 was updated to exclude HSAs with fewer than 500 covered lives from the calculation of the share of HSA-level variation in spending that was attributable to HRRs. The updated table is shown below.

Given that the IOM does not intend to release an updated version of PHE's report, we have focused here on outstanding substantive issues. Specifically, RTI has raised a concern about how PHE incorporated Medicaid managed care spending into its total spending measure. To calculate Medicaid managed care spending at the HRR level, PHE multiplied Medicaid fee-for-service spending at the HRR level by the ratio of Medicaid managed care spending per capita to Medicaid fee-for-service spending per capita at the state level.

PHE addressed outliers in the ratio by setting values below the 10th percentile or above the 90th percentile equal to the mean. RTI has argued that the high-value outliers may be valid, and recommended that PHE not set these values to the mean.

RTI's argument has some merit. On the other hand, trimming or winsorizing the upper end of the distribution of health spending or costs is common.[1] In general, such "data cleaning" can decrease or increase the bias arising from measurement error.[2] In PHE's view, neither PHE's nor RTI's approach is unambiguously superior.

In any event, PHE has calculated total spending in accordance with RTI's suggestion, and created alternative versions of Tables 15 through 19, reported below. The results of the analyses are qualitatively unchanged.

References

1. Wagstaff, A. and M. Lindelow, Can insurance increase financial risk?: The curious case of health insurance in China. Journal of Health Economics, 2008. 27(4): p. 990-1005.

2. Bollinger, C. and A. Chandra, Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data. Journal of Labor Economics, 2005. 23(2): p. 235-258.

PRECISION HEALTH ECONOMICS

Updated Table 13. Share of Variation in HSAs that is Attributable to HRRs

	MarketScan	OptumInsight*	Medicare	Medicaid
Total Spending	0.58	0.40	0.55	0.41
Input Price Adjusted Spending	0.47	0.37	0.59	0.42
Inpatient Admissions	0.28	0.49	0.38	0.56
Outpatient Visits	0.58	0.67	0.48	0.71
Rx Fills	0.62	0.42	0.44	0.74
ED Visit Days	0.42	0.37	0.30	0.67
Imaging Encounters	0.57	0.59	0.47	0.76

Alternative Table 15. Summary of Constructed Total Spending Measure

	Total Spending (Unadjusted)	Total Spending (Input Price Adjusted)
Mean	$522	$521
Standard Deviation	102.15	103.12
C.V.	0.20	0.20

Notes: These findings reflect the "control" specification which includes only predictors for year dummies and partial year enrollment.

Alternative Table 16. Correlation Between Total Spending and Population-Specific Spending

	MarketScan	Medicare	Medicaid
Total Spending (Input Price Adjusted)	0.17	0.27	0.65

Notes: These findings reflect the control specification which includes only predictors for year dummies and partial year enrollment.

PRECISION HEALTH ECONOMICS

Alternative Table 17. Impact of Additional Predictors on CV for Input Price Adjusted Total Spending

	Control	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 7	Cluster 9
C.V.	0.198	0.156	0.152	0.156	0.151	0.148	0.142

Alternative Table 18. R-squared Values of Medicare Quality Regression Specifications

	IQI	PQI	PSI	Other Predictors?	Predictor of Interest
Specification 1	0.01	0.11	0.01	N	Total
Specification 2	0.07	0.11	0.04	N	Medicare
Specification 3	0.53	0.73	0.57	Y	Total
Specification 4	0.55	0.76	0.57	Y	Medicare

Alternative Table 19. Standardized Coefficient on Independent Variable of Interest across Specifications

	Medicare Spending		Total Spending	
IQI Composite	**-0.26**	**-0.27**	0.10	-0.09
PQI Composite	**0.33**	**0.34**	**0.34**	0.11
PSI Composite	**0.20**	0.01	0.10	0.06
Specification	2	4	1	3
Other Predictors?	N	Y	N	Y

Notes: Predictors include Age, Sex, Age*Sex, PYE, Health Status, Race, Income, specialists/1,000; beds/1,000; HHI bed; % HMO; %Uninsured; Total Pop; Teaching Hospital; Malpractice GPCI
Values in **bold** are statistically significant at the 5% level.