Published in final edited form as:
N Engl J Med. 2019 April 18; 380(16): 1546–1554. doi:10.1056/NEJMsa1807379.

Accuracy of Valuations of Surgical Procedures in the Medicare Fee Schedule

David C. Chan, M.D., Ph.D., **Johnny Huynh, B.A.**, and **David M. Studdert, LL.B., Sc.D., M.P.H.**
Center for Health Policy–Center for Primary Care and Outcomes Research, Stanford University School of Medicine (D.C.C., D.M.S.), and Stanford Law School (D.M.S.), Stanford, the Department of Medicine, Veterans Affairs Palo Alto Health Care System, Palo Alto (D.C.C.), and the Department of Economics, University of California Los Angeles, Los Angeles (J.H.) — all in California.

Abstract

BACKGROUND—The Relative Value Scale Update Committee (RUC) of the American Medical Association plays a central role in determining physician reimbursement. The RUC's role and performance have been criticized but subjected to little empirical evaluation.

METHODS—We analyzed the accuracy of valuations of 293 common surgical procedures from 2005 through 2015. We compared the RUC's estimates of procedure time with "benchmark" times for the same procedures derived from the clinical registry maintained by the American College of Surgeons National Surgical Quality Improvement Program (NSQIP). We characterized inaccuracies, quantified their effect on physician revenue, and examined whether re-review corrected them.

RESULTS—At the time of 108 RUC reviews, the mean absolute discrepancy between RUC time estimates and benchmark times was 18.5 minutes, or 19.8% of the RUC time. However, RUC time estimates were neither systematically shorter nor longer than benchmark times overall (β, 0.97; 95% confidence interval, 0.94 to 1.01; P = 0.10). Our analyses suggest that whereas orthopedic surgeons and urologists received higher payments than they would have if benchmark times had been used ($160 million and $40 million more, respectively, in Medicare reimbursement in 2011 through 2015), cardiothoracic surgeons, neurosurgeons, and vascular surgeons received lower payments ($130 million, $60 million, and $30 million less, respectively). The accuracy of RUC time estimates improved in 47% of RUC revaluations, worsened in 27%, and was unchanged in 25%. (Percentages do not sum to 100 because of rounding.)

CONCLUSIONS—In this analysis of frequently conducted operations, we found substantial absolute discrepancies between intraoperative times as estimated by the RUC and the times recorded for the same procedures in a surgical registry, but the RUC did not systematically overestimate or underestimate times. (Funded by the National Institutes of Health.)

In 2017, The Medicare program made $70 billion in fee-for-service payments to physicians.[1] Payment levels for each type of service are determined by a formula that combines several

Address reprint requests to Dr. Chan at the Center for Health Policy/PCOR, 117 Encina Commons, Stanford CA, 94305, or at david.c.chan@stanford.edu.

elements — the nature of the work, practice expenses, the costs of malpractice insurance, and geographic differences in price — to arrive at a total number of relative value units (RVUs).[2] RVUs form the backbone of the Medicare Physician Fee Schedule, which many government programs and private insurers use to determine physician payments.

The work element, known as the "work RVU," determines approximately half of the total amount Medicare pays for physician services.[2] The work RVU is intended to reflect both the time it takes to perform a service and the intensity of the service (a function of the mental effort, technical skill, and stress involved in service delivery).[2,3] Time is a particularly influential factor.[4–6] For example, the service times used in valuing the sample of surgical procedures in this study explained 81% of the variance in the work RVUs assigned to those procedures (Fig. S1 in the Supplementary Appendix, available with the full text of this article at NEJM.org).

The Centers for Medicare and Medicaid Services (CMS) is legally responsible for setting and updating work RVUs. In practice, these tasks are largely delegated to the Relative Value Scale Update Committee (RUC), a group of physicians convened by the American Medical Association. The rigor and accuracy of the RUC's decisions have been questioned repeatedly.[7–18] One recurrent concern relates to the accuracy of measures of service time and the RUC's reliance on small-scale physician surveys to estimate that time.[5,6,11–13,19–22] We investigated the accuracy of the measures of time the RUC used to determine work RVUs for 293 common surgical procedures.

Methods

The RUC and Work RVU Determinations

The RUC currently has 31 members, 25 of whom are appointed by the major specialty societies. Approximately half of all RUC reviews are valuations of new or modified services; the rest involve revaluations of existing services.

Until 2012, there were two main types of revaluation: ad hoc reviews ("annual reviews") and 5-year reviews.[2,3,12] The 5-year reviews, which were discontinued after 2012, involved comprehensive reviews of potentially misvalued services in the fee schedule. This program identified services for review mainly through a public comment process administered by CMS.[3] In annual reviews, however, the RUC exercises greater control and substantial discretion over which services are selected for review.

Specialty societies with relevant expertise lead reviews. They develop and advocate for a work RVU level, using results from a survey to justify their recommendation.[2,9,12,23] The survey describes the service and presents a vignette that depicts delivery to a typical patient. Respondents estimate the time needed to render the service and rate its complexity. Specialty societies must administer the survey to a sample of at least 30 practicing physicians who are familiar with the service.

Recommendations approved by a two-thirds majority of voting members are submitted to CMS, which has historically accepted more than 90% of RUC recommendations.[3,24]

Additional details regarding the RUC and its review procedures are provided in Section 2 in the Supplementary Appendix.

Sources of Data

We obtained data from three sources: the RUC, the American College of Surgeons National Surgical Quality Improvement Program (NSQIP), and CMS.

RUC—The RUC provided us with records dating from its inaugural meeting in May 1992 through 2015. The data included Current Procedural Terminology (CPT) codes of reviewed services, review outcomes, and estimates of service time from the physician surveys that were used in determining work RVUs. For surgical procedures, values for the service-time variable indicated the estimated number of minutes from incision to closure in a typical case.

Work RVUs for surgical procedures also encompass certain tasks rendered before and after the operation. The RUC provided data on these tasks, which enabled us to isolate the incisionto-closure component of each RVU. Our analyses focused on that component.

NSQIP—The NSQIP, established in the mid-1980s, has grown to encompass more than 600 participating hospitals that contribute data on surgical cases to a central clinical registry.[25,26] The NSQIP's sampling strategy, data abstraction procedures, and database are described elsewhere.[27] We obtained the NSQIP Participant Use Data File for 2005 through 2015. The file contains many caselevel variables, including the CPT code and the number of minutes from incision to closure.

CMS—The Physician/Supplier Procedure Summary (PSPS) files are an annual summary of information on all claims for services and durable medical equipment submitted to Medicare's Part B fee-for-service program.[28] Annual counts of approved claims are aggregated at several levels, including CPT code and physician specialty. We obtained PSPS files for the period 2011 through 2015.

Study Sample

We used NSQIP data to identify the 300 most commonly performed surgical procedures. We then linked each procedure to its RUC intraoperative time estimates, NSQIP intraoperative times, and annual volume and charges from the PSPS files. RUC time estimates were missing for 7 procedures, leaving 293 procedures in our final study sample. These 293 procedures accounted for 3.9 million cases in the 2005–2015 NSQIP data (mean, 13,317 cases per procedure; range, 2236 to 217,447), or 85% of all cases recorded in the NSQIP database during this period.

Statistical Analysis

We created cross-sectional and longitudinal measures to assess the accuracy of the RUC estimates of procedure time.

Accuracy at Review—To derive the benchmark time for a procedure, we gathered intraoperative times for all reports of the procedure recorded in the NSQIP database in the two calendar quarters before and the two calendar quarters after the RUC review. Because this approach required contemporaneity in the RUC and NSQIP measures, it had to be restricted to 108 RUC reviews of 98 sampled procedures performed in the period covered by our NSQIP data (2005–2015).

We conceived of accuracy as the difference, or discrepancy, in minutes between the RUC estimate and the mean value of its NSQIP times. (In sensitivity analyses, we also examined discrepancies calculated on the basis of differences from the median value of NSQIP times.) Discrepancies were characterized in two ways. First, we calculated the mean absolute discrepancy, an average of the size of the discrepancies between the RUC and NSQIP times that disregards their directionality. Second, we conducted a bivariate linear regression analysis, weighting the analysis according to procedure frequency and excluding an intercept term. A value of 1 on the coefficient estimated in this analysis would indicate that on average, across procedures, there was no systematic difference between RUC and NSQIP times. Deviation from 1 indicates the extent to which RUC time estimates are systematically longer (>1) or shorter (<1) than NSQIP times.

Longitudinal and Overall Accuracy—We used the same sample of procedures and a similar approach to derive longitudinal measures of accuracy, except that benchmark values came from a moving four-quarter window. When repeated and averaged over the period in which a procedure’s work RVU was in force, these calculations produced a measure (in minutes) of total discrepancy. Thus, total discrepancy is a function of two components: first, the initial discrepancy, established at the moment of RUC review, and second, the temporal changes in the benchmark times, which may have exacerbated or corrected the initial discrepancy. We calculated the contribution of each component to total discrepancy. (See Section 3 in the Supplementary Appendix for additional details.)

Effects on Revenue—We used the PSPS file to calculate actual Medicare payments from 2011 through 2015 for each of the 293 procedures in our sample and compared these payments with an estimate of the counterfactual amount Medicare would have paid physicians if the work RVU had been valued with the use of up-to-date times from the NSQIP data instead of the RUC time estimates. “Gains” and “losses” were the difference between the estimate of counterfactual work revenue and actual work revenue. We stratified these estimates according to physician specialty.

Our analysis assumed that the incision-to-closure (“intraservice”) component of the work RVU would scale proportionally with any changes in the intraservice time — in other words, that dimensions of work other than procedure time (“intraservice work per unit of time”) were constant.[29] (Details of the analysis of the effects on revenue, including justification for our scaling assumption, are provided in Section 4 in the Supplementary Appendix.)

Selection of Procedures for RUC Review—We used logistic-regression analysis to estimate the relationship between the size of discrepancies in time and the probability that a procedure would be selected for RUC review. This analysis was based on a procedure–RUC meeting-level data set that consisted of 8071 observations (293 procedures in up to 31 meetings per procedure). The outcome variable specified whether a procedure was reviewed at a meeting. The independent variable of interest was the discrepancy in time (positive or negative) for the procedure in the year of the meeting. (Model details are provided in Section 5 in the Supplementary Appendix.)

Effect of RUC Re-Reviews on Accuracy of Time Estimates for Procedures—To determine how RUC re-reviews affected discrepancies, we compared three values: the time estimate on which the prevailing work RVU was based, the time estimate RUC produced for the new review, and the mean NSQIP time in the four-quarter window surrounding the new review. This analysis was confined to 102 RUC reviews conducted in 2005 through 2015 for which all three time values were available.

Results

Characteristics of the Study Sample

Between January 1, 2005, and December 31, 2015, the RUC held 33 meetings (3 per year) at which it conducted 2717 reviews of 2492 distinct services (mean, 82.3 services per meeting; median, 63). A total of 108 of these reviews addressed 98 of the procedures in our sample. On average, each procedure in our sample had a 1.3% chance of being reviewed at a given meeting.

The RUC time estimates for procedures in our sample ranged from 15 minutes for CPT 10140 (Incision and Drainage on Skin) to 306 minutes for CPT 35566 (Vein Bypass Graft), with an average of 114 minutes per procedure. The physician surveys that produced these time estimates had a mean of 58 respondents (median, 53; interquartile range, 38 to 73) and a mean response rate of 21%. (A complete list of the sampled procedures and their RUC and NSQIP time values are provided in Section 8 in the Supplementary Appendix.)

Accuracy at Review

In 108 reviews of sampled procedures, 57 reviews relied on RUC time estimates that were shorter than the benchmark times (mean underestimate, 23 minutes) and 51 reviews relied on RUC time estimates that were longer than the benchmark times (mean overestimate, 14 minutes) (Fig. 1). The mean absolute discrepancy was 18.5 minutes, or 19.8% of the RUC time. However, RUC time estimates were neither systematically shorter nor systematically longer than benchmark times: the regression coefficient indicated that they were 3% shorter, but this small difference was not significant (β, 0.97; 95% confidence interval [CI], 0.94 to 1.01; P = 0.10).

Longitudinal Accuracy

Figure 2 shows two components of discrepancy for each of the 98 procedures. The first component is the initial discrepancy at the time of review — the same form of inaccuracy indicated by deviations from the 45-degree line in Figure 1. The second component reflects temporal changes in the benchmark time that occurred after review. Figure 2 also shows the total discrepancy for each procedure over the period of observation, after temporal changes in procedure time were added to or subtracted from the initial discrepancy. Temporal changes exacerbated the initial discrepancy for some procedures and partially corrected it for others. When combined across all procedures, these exacerbations and corrections were roughly equal in size.

On average, the size of total discrepancies was 18% of the benchmark times, although the extent varied widely according to procedure (range, 2 to 58%). There was a substantial amount of temporal change in the duration of many procedures; the size of absolute temporal change was 43% of the average size of the initial discrepancy. Across all procedures, exacerbations and corrections were roughly equal in size.

Effects of Discrepancies in Procedure Times on Physician Revenue

Figure 3 shows the estimated differences between actual Medicare payments to physicians between 2011 and 2015 for performing the 293 procedures in our sample and the estimated amounts Medicare would have paid if the work RVU had been based on the up-to-date benchmark time rather than the RUC's estimate of time. Differences are presented for each of the eight surgical specialties to which Medicare paid the most for these procedures.

The estimates suggest overpayments to surgeons in two of the specialties, underpayments to surgeons in five of the specialties, and virtually no difference in payments to general surgeons. Orthopedic surgeons received payments of $160 million more than they would have received had benchmark times been used instead of the RUC time estimates, an amount that represents 1.7% of total Medicare Part B payments made to these surgeons during the 5-year period. Urologists received $40 million in overpayments (1.0% of total payments made to them). However, surgeons in several specialties received substantial underpayments. Cardiothoracic surgeons received $130 million (6.4%) less than they would have received if benchmark times had been used, neurosurgeons $60 million (2.9%) less, vascular surgeons $30 million (1.5%) less, plastic surgeons $20 million (1.8%) less, and obstetrician-gynecologists $20 million (1.1%) less. (Fig. S6 in the Supplementary Appendix shows the effects on revenue for the average surgeon in each of these specialties.)

Relationship between Discrepancy in Procedure Time and Selection for RUC Review

Among annual reviews, there was a positive association between the size of discrepancy in procedure time and the probability that it would be selected for RUC review (Fig. S7A in the Supplementary Appendix). Procedures with little or no discrepancy had an approximate chance of 0.6% of being selected for review at any given RUC meeting, whereas procedures with discrepancies of 75 minutes or more had an approximate chance of 1% of being selected. The association was similar for positive and negative discrepancies. In other words, procedures that took longer than the RUC-estimated time were not appreciably more likely to be selected for review than procedures that were shorter than the RUC-estimated time. Among procedures selected for 5-year reviews (a selection process over which the RUC had less control), there was no clear association between the time discrepancy and the probability of RUC review (Fig. 7B in the Supplementary Appendix).

Effect of RUC Re-Reviews on Discrepancies

In 2005 through 2015, the RUC conducted rereviews of 48% (140 of 293) of the procedures in our sample, of which 102 had the requisite NSQIP benchmark data required for inclusion in the analysis of re-reviews (Fig. 4). Nearly half the re-reviews (48 of 102) reduced discrepancies, shifting the new RUC time estimate closer to the benchmark time. However, in 28 re-reviews the new RUC time estimate shifted further away from the benchmark time, and in 26 re-reviews the discrepancy was unchanged. Overall, the average absolute discrepancy immediately before re-review was 21% of the benchmark time; immediately after re-review, the discrepancy decreased to 15% of the benchmark time.

Discussion

In this study, we analyzed estimates of the service times used to set reimbursement levels in the Medicare Physician Fee Schedule for the most commonly performed surgical procedures. By comparing RUC time estimates with times for the same procedures listed in a large registry of surgical times, we found discrepancies. These discrepancies varied across procedures in both size and direction, but we found no evidence that the RUC systematically overestimated or underestimated procedure times. For many procedures, the discrepancies changed substantially in size over time. Our analyses suggest that these inaccuracies had nontrivial distributional effects on surgeons' income from clinical practice. Rereviews by the RUC helped to reduce the inaccuracies. However, only half the procedures in our sample underwent re-review in the years 2005 through 2015, and half of those re-reviews did not reduce inaccuracies.

Recognizing the powerful influence that the RUC's time estimates exert on prices in the Medicare Physician Fee Schedule, the authors of several previous studies[5,19–21] have investigated their accuracy and potential bias. A consistent finding is that the RUC's survey-derived measures frequently exceed the service times in estimates from other sources. These studies have design weaknesses: some examined relatively few services,[19–21] one drew data from only two clinics,[19] one imputed benchmark times rather than measuring them directly,[5] and none sought to quantify the effects of inaccuracies on physician reimbursement.

However, the most striking limitation of the existing research — and a key motivation for our study — is its crude treatment of chronological time. Previous studies did not attempt to measure benchmark times at the same time that the RUC surveys were conducted, creating a legitimate reason for at least some discrepancy. More important, these studies focused on cross-sectional accuracy, largely ignoring longitudinal dimensions of accuracy. The size of the discrepancy between a procedure's RUC time estimate and its benchmark time typically expands or contracts over time, so effects cannot be fully understood without a longitudinal perspective.

Our finding that the RUC did not systematically overestimate procedure times conflicts with results from previous studies. The conflict may be partially explained by the fact that earlier studies used median time values in defining benchmarks, whereas we used mean values. When we modified our analysis to benchmark against median NSQIP times, the RUC time estimates were 9% longer than the benchmark times at the time of review (β, 1.09; 95% CI, 1.06 to 1.13; P<0.001) (Fig. S11 in the Supplementary Appendix); this difference is significant but substantially smaller than most previous estimates. (For a discussion of the nature and merits of alternative measures of inaccuracy, see Section 7 in the Supplementary Appendix.)

The absence of systematic divergence from benchmark times is somewhat reassuring: it suggests that overall, inaccurate time estimates run in both directions and more or less cancel each other out. At the level of physicians, specialties, and hospitals, however, distortions may be systematic and large. Whether their net effect is financially beneficial or detrimental will vary according to the size and direction of discrepancies in the mix of procedures involved. Our revenue analysis shows this variation at the specialty level.

Our findings point to two reforms that have the potential to improve the accuracy of service valuations. First, the RUC may benefit from larger and more reliable sources of data for the time estimates it uses in determining work RVUs; we are not the first to make this recommendation.[6,19,21] The RUC has sometimes obtained and considered service-time data from clinical registries, including NSQIP data and the Society of Thoracic Surgeons national database,[30] but this practice is uncommon. Second, the realtime accuracy of the RUC's valuations could be enhanced by monitoring such data sources for substantial changes in the duration of procedures and using this information to prioritize procedures for re-review.

Our study has several limitations. First, by focusing on the accuracy of the RUC's time estimates for surgical services, we did not analyze medical services, nor did we consider what effects a schedulewide recalibration of service times would have. Second, we did not examine other components of RVUs, such as work-intensity or practice-expense RVUs. Our analysis of revenue effects assumed that changes in procedure time did not affect procedure intensity, which is consistent with the concept of fixed intraservice work per unit of time.[29]

Finally, although the NSQIP database draws data from hundreds of hospitals and is the largest repository of detailed data on surgical cases, contributing hospitals are not nationally representative,[31] a fact that may have skewed the procedure times we used as benchmarks. We explored this possibility and did not detect evidence that our findings were substantially affected by nonrepresentativeness in NSQIP data. (Details of these sensitivity analyses are provided in Section 6 in the Supplementary Appendix.)

The federal government is changing aspects of how it pays physicians.[32–34] However, the approach used by CMS since 1992 to value physician services remains fundamentally intact. Reforms to that approach may improve the accuracy and fairness of reimbursement. Whether it is feasible to implement the necessary reforms within the existing institutional and procedural framework is a question that warrants careful consideration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by a Director's Early Independence Award (1DP5OD019903–01, to Dr. Chan) from the National Institutes of Health.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Atul Gawande, Miriam Laugesen, Ateev Mehrotra, Joseph Newhouse, Sherry Smith, and Peter K. Smith, who provided helpful comments on an earlier draft of the manuscript.

References

1. Congressional Budget Office. Baseline projections for Medicare. June 29, 2017 (https://www.cbo.gov/sites/default/files/recurringdata/51302-2017-06-medicare.pdf).
2. Medicare RBRVS 2018: the physicians' guide. Chicago: American Medical Association, 2018.
3. American Medical Association. RVS update process. 2018 (https://www.ama-assn.org/sites/default/files/media-browser/public/rbrvs/ruc-update-booklet_0.pdf).
4. Wynn BO, Burgette LF, Mulcahy AW, et al. Development of a model for the validation of work relative value units for the Medicare physician fee schedule Santa Monica, CA: RAND, 2015 (https://www.rand.org/pubs/research_reports/RR662.html).
5. Burgette LF, Mulcahy AW, Mehrotra A, Ruder T, Wynn BO. Estimating surgical procedure times using anesthesia billing data and operating room records. Health Serv Res 2017; 52: 74–92. [PubMed: 26952688]
6. Zuckerman S, Merrell K, Berenson R, Mitchell S, Upadhyay D, Lewis R. Collecting empirical physician time data: piloting an approach for validating work relative value units. Washington, DC: Urban Institute, 2016 (https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/Downloads/Collecting-Empirical-Physician-Time-Data-Urban-Report.pdf).
7. Ginsburg PB, Berenson RA. Revising Medicare's physician fee schedule — much activity, little change. N Engl J Med 2007; 356: 1201–3. [PubMed: 17377156]
8. Goodson JD. Unintended consequences of resource-based relative value scale reimbursement. JAMA 2007; 298: 2308–10. [PubMed: 18029836]
9. Medicare Payment Advisory Commission. Medicare payment policy: report to the Congress. March 2006 (http://www.medpac.gov/docs/default-source/reports/Mar06_EntireReport.pdf).
10. Matthews AW, McGinty T. Physician panel prescribes the fees paid by Medicare. Wall Street Journal. October 26, 2010.
11. Whoriskey P, Keating D. How a secretive panel uses data that distorts doctors' pay. Washington Post. July 20, 2013.
12. Laugesen MJ. Fixing medical prices: how physicians are paid. Cambridge, MA: Harvard University Press, 2016.
13. Reinhardt UE. The little-known decision-makers for Medicare physician fees. New York Times. December 10, 2010 (https://economix.blogs.nytimes.com/2010/12/10/the-little-known-decision-makers-for-medicare-physicans-fees/).
14. Bodenheimer T, Berenson RA, Rudolf P. The primary care-specialty income gap: why it matters. Ann Intern Med 2007; 146: 301–6. [PubMed: 17310054]
15. Berwick Fischer v., 2012 U.S. Dist. LEXIS 65034 (D. Md. May 9, 2012).
16. Klepper B The RUC, health care finance's star chamber, remains untouchable. Health Affairs Blog. February 1, 2013 (https://www.healthaffairs.org/do/10.1377/hblog20130201.027753/full/).
17. Medicare Payment Advisory Commission. Medicare payment policy: report to the Congress. March 2008 (http://www.medpac.gov/docs/default-source/reports/mar08_entirereport.pdf?sfvrsn=0).
18. Replace the RUC! home page (https://replacetheruc.net/).
19. McCall N, Cromwell J, Braun P. Validation of physician survey estimates of surgical time using operating room logs. Med Care Res Rev 2006; 63: 764–77. [PubMed: 17099125]
20. Cromwell J, McCall N, Dalton K, Braun P. Missing productivity gains in the Medicare physician fee schedule: where are they? Med Care Res Rev 2010; 67: 676–93. [PubMed: 20555013]
21. Braun P, McCall N. Methodological concerns with the Medicare RBRVS payment system and recommendations for additional study. Research Triangle Park, NC: RTI International, 2011 (http://67.59.137.244/documents/Aug11_Methodology_RBRVS_contractor.pdf).

22. Eaton J Little-known AMA group has big influence on Medicare payments. Kaiser Health News. October 27, 2010 (https://khn.org/news/ama-center-public-integrity/).

23. American Medical Association. RVS Update Committee (RUC) (https://www.ama-assn.org/rvs-update-committee-ruc).

24. Laugesen MJ, Wada R, Chen EM. In setting doctors' Medicare fees, CMS almost always accepts the Relative Value Update Panel's advice on work values. Health Aff (Millwood) 2012; 31: 965–72. [PubMed: 22566435]

25. American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP). ACS NSQIP: how it works (http://site.acsnsqip.org/wp-content/uploads/2012/02/TechnicalPaper1.pdf).

26. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. J Am Coll Surg 2013; 217(2): 336–46.e1. [PubMed: 23628227]

27. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-con-trolled program for the measurement and enhancement of the quality of surgical care. Ann Surg 1998; 228: 491–507. [PubMed: 9790339]

28. Centers for Medicare & Medicaid Services. Physician/supplier procedure summary (https://www.cms.gov/research-statistics-data-and-systems/files-for-order/nonidentifiabledatafiles/physiciansupplierproceduresummarymasterfile.html).

29. Mabry CD, McCann BC, Harris JA, et al. The use of intraservice work per unit of time (IWPUT) and the building block method (BBM) for the calculation of surgical work. Ann Surg 2005; 241: 929–40. [PubMed: 15912042]

30. The Society of Thoracic Surgeons. STS national database (https://www.sts.org/registries-research-center/sts-national-database).

31. Sheils CR, Dahlke AR, Kreutzer L, Bilimoria KY, Yang AD. Evaluation of hospitals participating in the American College of Surgeons National Surgical Quality Improvement Program. Surgery 2016; 160: 1182–8. [PubMed: 27302100]

32. Medicare Access and CHIP Reauthorization Act of 2015 (MACRA), Pub. L. No. 114–10.

33. Burwell SM. Setting value-based payment goals — HHS efforts to improve U.S. health care. N Engl J Med 2015; 372: 897–9. [PubMed: 25622024]

34. Centers for Medicare & Medicaid Services (CMS). Medicare program: Merit-Based Incentive Payment System (MIPS) and Alternative Payment Model (APM) incentive under the Physician Fee Schedule, and criteria for physician-focused payment models: final rule with comment period. Fed Regist 2016; 81(214): 77008–831. [PubMed: 27905815]

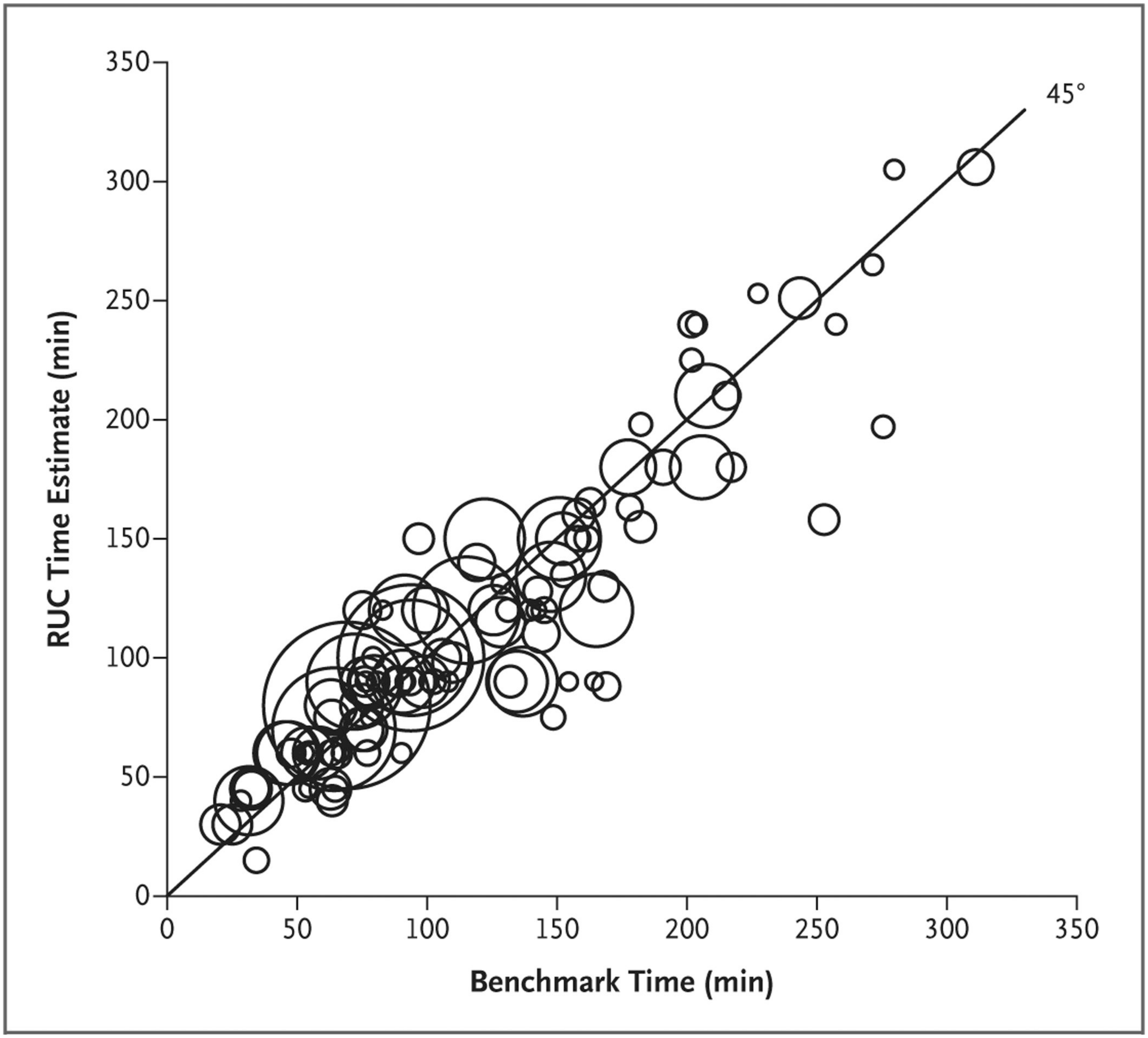

Figure 1. Relationship between RUC Time Estimates and Benchmark Times in 108 Reviews of Common Surgical Procedures.

Each bubble represents a surgical procedure reviewed by the Relative Value Scale Update Committee (RUC) from 2005 through 2015. Bubble size represents the frequency with which the procedures were performed, as recorded in the clinical registry maintained by the American College of Surgeons National Surgical Quality Improvement Program (NSQIP). Benchmark times were defined as mean times in the NSQIP data for the relevant procedure and calendar quarter. The 45-degree line indicates equivalence between the RUC time and the benchmark time, and the distance of bubbles from this line indicates the degree of discrepancy between the two time values.

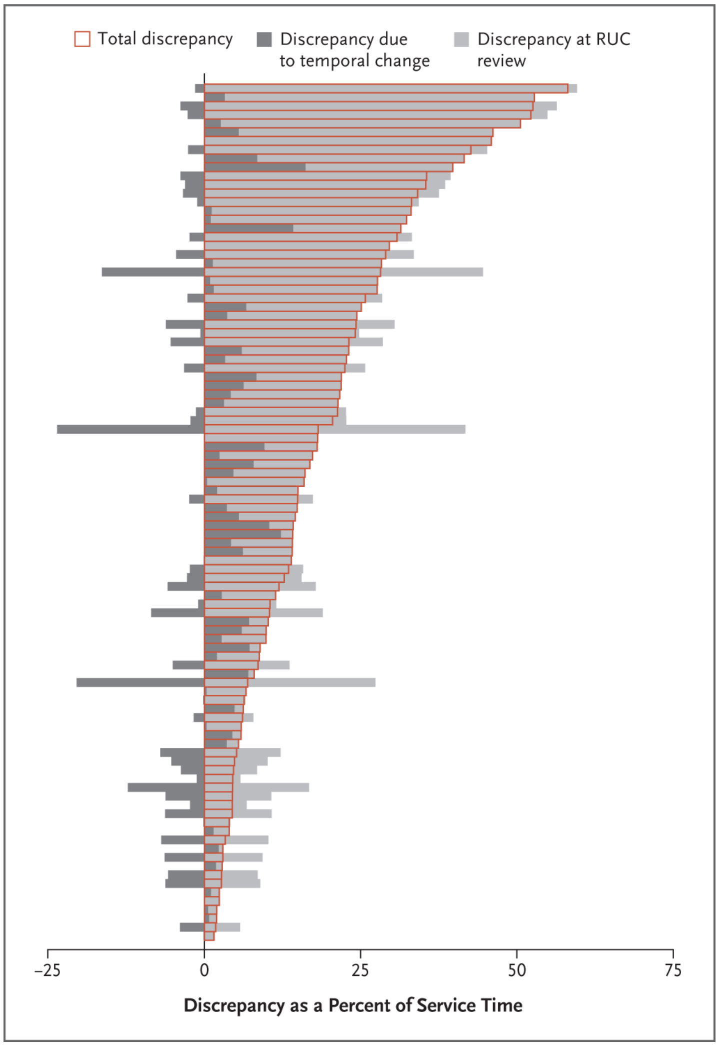

Figure 2. Discrepancies between RUC Time Estimates and Benchmark Times for 98 Common Surgical Procedures.

Shown are the components of discrepancy for each of 98 procedures performed, expressed as an average percentage deviation from the procedure's benchmark time over the period of observation. Each bar represents a procedure. The light gray segment of each bar represents the initial discrepancy at the time of review. The dark gray segments represent temporal changes in procedure times after review. Dark gray segments to the left of the zero line indicate temporal changes that decreased discrepancy from the initial discrepancy; dark gray segments to the right of the zero line indicate temporal changes that increased discrepancy from the initial discrepancy. The boxes outlined in red indicate the total discrepancy after adding temporal changes to or subtracting them from the initial discrepancy.

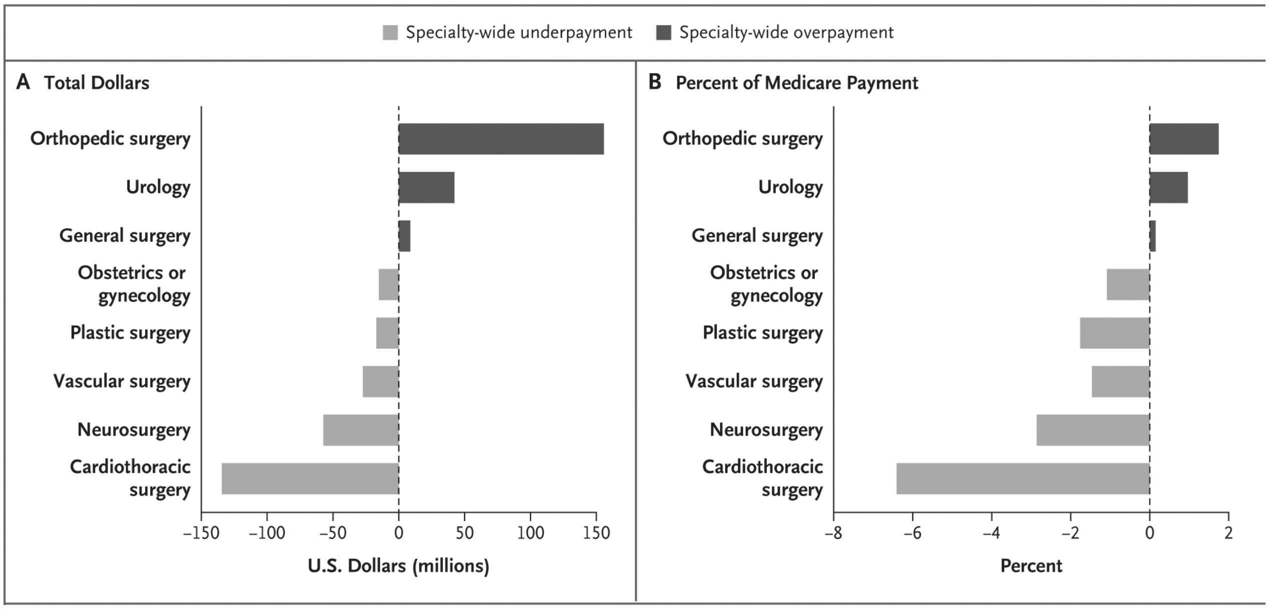

Figure 3. Effect of Inaccuracies in Time Estimates for 293 Procedures on Medicare Reimbursement for Eight Surgical Specialties, 2011–2015.

The bars show the differences, according to specialty, between actual Medicare payments and an estimate of counterfactual Medicare payments had the procedures been valued according to benchmark (NSQIP) times instead of RUC time estimates. Panel A shows the differences in dollars, and Panel B the differences as a percentage of all Medicare payments (i.e., for all services) made to physicians in the specialties. The dark gray bars to the right of the zero line indicate specialty-wide overpayment relative to the counterfactual amount. Light gray bars indicate specialty-wide underpayment.

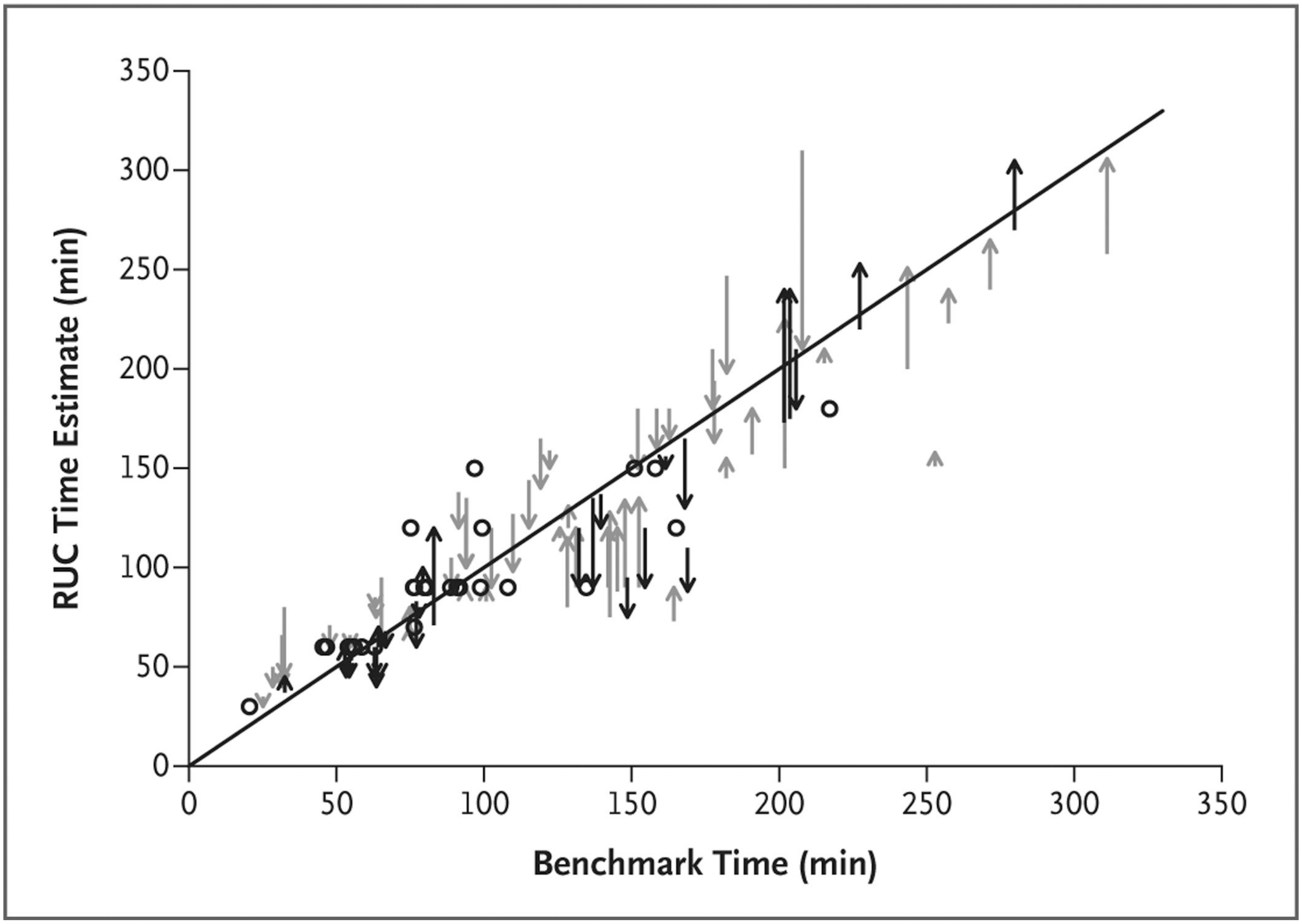

Figure 4. Effect of RUC Re-Reviews on Extent of Discrepancies in Procedure Time, 2005–2015.

Each arrow and each dot represents a RUC re-review. The distance between the base of each arrow and the 45-degree line represents the size of the discrepancy between the RUC time measure used in the prevailing work RVU and the benchmark time (mean NSQIP) immediately before re-review. The distance between the tip of each arrow and the 45-degree line represents the size of the discrepancy between the new RUC time estimate used in the re-review and the benchmark time. The dots indicate re-reviews in which the discrepancy between the new RUC time and the benchmark time did not change. Six re-reviews conducted in 2005 through 2015 were excluded because their initial reviews were not conducted by the RUC and the time measures used in those reviews were not in our data.