Research Report

Medicare Imaging Demonstration Final Evaluation

Report to Congress

Justin W. Timbie, Peter S. Hussey, Lane F. Burgette, Neil S. Wenger, Afshin Rastegar, Ian Brantley, Dmitry Khodyakov, Kristin J. Leuschner, Beverly A. Weidmer, Katherine L. Kahn

Sponsored by the Centers for Medicare & Medicaid Services

For more information on this publication, visit www.rand.org/t/rr706

Published by the RAND Corporation, Santa Monica, Calif.

Preface

Increasing use of advanced medical imaging is often cited as a key driver of cost growth in medical spending. In 2011, the Medicare Imaging Demonstration (MID) from the Centers for Medicare & Medicaid Services (CMS) began testing whether exposing ordering clinicians to appropriateness guidelines for advanced imaging would reduce or eliminate ordering of inappropriate advanced images. Decision support systems (DSSs) were selected as the vehicle for delivering these guidelines for Medicare fee-for-service patients. The DSS tool is intended to provide immediate physician feedback on the appropriateness of a test ordered for a patient, based on current medical specialty guidelines.

The MID was authorized by Section 135(b) of the Medicare Improvements for Patients and Providers Act of 2008. The statute instructs the U.S. Department of Health and Human Services (HHS) to collect data on fee-for-service Medicare patients to determine the appropriateness of services in relation to established criteria and to provide participating physicians with feedback reports that permit comparison against physician peers on adherence to appropriateness criteria. The statute emphasizes the use of DSSs to achieve these goals and prohibits the use of prior authorization requirements. CMS, an agency within HHS, entered into a four-year contract with the RAND Corporation in 2010 to evaluate the demonstration. The overall goal of this project is to determine the extent to which exposure to medical specialty guidelines through DSSs is associated with changes in the use of advanced imaging, if at all.

This report, written by RAND, describes results from RAND's mixed-methods evaluation using combinations of quantitative and qualitative analyses of decision support data, Medicare claims, focus groups, and surveys. Based on these analyses, we also provide recommendations for the consideration of policymakers. An executive summary of RAND's report, written by CMS and submitted as the MID Report to Congress with RAND's report, is also included.

This work was sponsored by CMS under contract No. HHSM-500-2005-00028I, for which David Nyweide served as the contracting officer's representative. The research was conducted in RAND Health, a division of the RAND Corporation. A profile of RAND Health, abstracts of its publications, and ordering information can be found at www.rand.org/health.

This page is intentionally blank.

Contents

This page is intentionally blank.

Figures

This page is intentionally blank.

Tables

Medicare Imaging Demonstration Evaluation Executive Summary[1]

Section 135(b) of the Medicare Improvements for Patients and Providers Act of 2008 (P.L. 110-275) (MIPPA) required the Secretary of Health and Human Services to conduct a demonstration project in which data regarding physician compliance with appropriateness criteria are collected to determine the appropriateness of advanced diagnostic imaging services furnished to Medicare beneficiaries. Designed as an alternative to prior authorization, the Medicare Imaging Demonstration (MID) informed physicians about the appropriateness of their orders according to appropriateness criteria selected by the Secretary and programmed into computer order-entry systems known as decision support systems (DSSs).

The evaluation of MID sought to quantify rates of appropriate, uncertain, and inappropriate advanced diagnostic image ordering in the Medicare program and to determine whether exposing physicians to guidelines at the time of order is associated with more appropriate ordering and an attendant change in utilization. Under section 135(b)(5) of MIPPA, the Secretary is required to evaluate the demonstration project and to submit a report to Congress containing the results of the evaluation and recommendations for legislation and administrative action, as the Secretary determines appropriate, no later than one year after completion of the demonstration.

The two-year demonstration launched October 1, 2011, for physicians in one of five participating conveners across geographically and organizationally diverse practice settings. A convener was a single entity responsible for providing and supporting the use of a DSS for a collection of physician practices. Participation in the demonstration was voluntary, and there were no negative payment consequences for billed services for not consulting DSS. The statute required the Secretary to reimburse physicians for reasonable administrative costs incurred in participating in the demonstration project and to provide reasonable incentives to physicians to encourage participation. To meet this requirement, the conveners and physician practices received payment for their costs of participation when they supplied DSS records for the advanced diagnostic imaging procedures furnished during the demonstration. Physician practices decided how to distribute their payments to physicians. While physicians were required to consult the DSS every time they ordered an advanced diagnostic imaging procedure, they retained the autonomy to continue with or change their orders after consulting DSS, with no financial incentive to order more or fewer imaging procedures.

The statute described three different types of models for collecting data on appropriateness of orders—a point of service model; a point of order model; and any other model that the Secretary determines to be useful in evaluating the use of appropriateness criteria for advanced diagnostic

[1] This executive summary of RAND's evaluation report was submitted by the Centers for Medicare & Medicaid Services as the MID Report to Congress. It was subject to RAND's Quality Assurance process.

imaging services. The demonstration tested the point of order model, as it reflected the state of the decision support market according to the environmental scan during the design phase of the demonstration. Any DSS could be used in the demonstration as long as it was programmed with the same appropriateness criteria used by all demonstration participants.

The contractor tasked with designing and operating the demonstration, the Lewin Group, identified the conditions and accompanying medical professional society guidelines associated with the 12 most common advanced diagnostic imaging procedures performed among Medicare beneficiaries—magnetic resonance imaging (MRI) of brain, knee, lumbar spine, or shoulder; computed tomography (CT) of abdomen, abdomen and pelvis, brain, lumbar spine, pelvis, sinus, or thorax; or Single Photon Emission Computed Tomography Myocardial Perfusion Imaging (SPECT MPI). When a physician—or a physician assistant or nurse practitioner who could legally order advanced diagnostic imaging—intended to order one of the 12 advanced diagnostic imaging procedures, he or she was required to consult a DSS programmed with the guidelines. The DSS, then, was intended to provide the ordering physician with instant feedback about the appropriateness of the order. If they entered a minimum of 30 rated orders over three- to six-month periods during the demonstration, physicians were also eligible to receive feedback reports about their appropriateness rates compared with the aggregated rates of their peers.

The data analyses and interpretation included in this report were prepared by the RAND Corporation (RAND) under contract with the Centers for Medicare & Medicaid Services (CMS). To perform its evaluation, RAND gathered information about the demonstration from the Lewin Group; analyzed DSS data and claims data; convened physician, staff, and patient focus groups that were supplemented with short questionnaires for physicians and staff; and conducted interviews with convener leadership. Most analyses were performed for each convener individually, rather than as a collective group, because the variety in structure of practices and DSSs fundamentally differentiated physicians' experience of the demonstration across conveners. To account for the impact of DSS on ordering behavior, 18 months of orders were analyzed relative to an initial 6-month baseline period of the demonstration during which orders were entered into DSS and rated without providing immediate appropriateness feedback to the orderer. To account for existing trends in utilization of advanced diagnostic imaging, analyses of any changes in utilization involved a matched comparison group that did not use decision support for Medicare patients' advanced diagnostic imaging orders.

In its evaluation report to CMS, RAND directly addressed the impact and implications of the demonstration (Medicare Imaging Demonstration Evaluation Report, Appendix A). This Report to Congress summarizes RAND's findings, including factors required to be assessed or analyzed under section 135(b)(5) of MIPPA.

Appropriate, Uncertain, and Inappropriate Ordering Rates and Patterns

In MID, advanced diagnostic imaging orders were entered into and rated by a DSS for appropriateness relative to four categories—"appropriate," "uncertain," "inappropriate," or "not covered by guidelines." "Appropriate" indicated that the order was consistent with the guidelines used in the demonstration. "Uncertain" meant that physicians should use their discretion because the guidelines for a given clinical scenario could not provide definitive guidance, while "inappropriate" signaled that the order was not consistent with guidelines. "Not covered by guidelines" displayed when the DSS order could not be linked to a guideline and, thus, could not be rated for appropriateness and was unrated in the demonstration. DSS orders could not be linked to a guideline when a physician's own reason for an order did not match the selected clinical indications in DSS used to link to a guideline or when a guideline simply does not exist for a given clinical scenario.

Over the course of the two-year demonstration, 3,916 physicians placed 139,757 initial orders for advanced diagnostic imaging procedures before receiving DSS feedback. Most physicians (70.5 percent of primary care physicians, 59.8 percent of medical specialists, and 64.6 percent of surgical specialists) placed fewer than 20 orders, or less than 1 order per month. A total of 8,345 orders (37.3 percent) during the baseline period and 40,536 orders (34.5 percent) during the intervention period could be rated for appropriateness (appropriate, uncertain, or inappropriate), resulting in a total of 48,881 (35.0 percent) rated orders that could be analyzed in both periods. The majority of orders could not be analyzed because they were "not covered by guidelines."

Among rated orders in the baseline period, between 61.5 percent and 81.8 percent were appropriate across conveners, representing the range of appropriate ordering rates in the fee-for-service Medicare program prior to exposing physicians to appropriateness criteria through DSS. Likewise, between 10.3 percent and 21.0 percent of rated orders were uncertain at baseline across conveners and 7.8 percent to 18.1 percent were inappropriate. Compared with the baseline period, all but one convener showed an increase in the rate of appropriate ordering —with decreases in the rates of uncertain and inappropriate ordering—for final rated orders after physicians received DSS feedback on their orders in the intervention period. Conveners ranged from 75.1 percent to 83.9 percent in their rates of appropriate ordering during the intervention period, with rates of uncertain ordering between 11.1 percent and 16.1 percent and rates of inappropriate ordering between 5.3 percent and 9.0 percent. While the conveners overall seemed to show an improvement in appropriate ordering between the baseline and intervention periods, the percentage of unrated orders varied over time as well. Therefore, if the orders "not covered by guidelines" could have been rated, they may have changed the percentage of appropriate, uncertain, and inappropriate orders. For this reason, these changes in rates do not necessarily indicate an improvement in the appropriate ordering rate over the course of the demonstration.

When including both rated and unrated orders to determine the proportion of appropriate, uncertain, inappropriate, and unrated orders, most conveners sustained stable levels of appropriate and inappropriate rated orders between the baseline and intervention periods of the demonstration. The only convener that exhibited relative improvements in appropriateness rates between periods showed an accompanying decrease in the rates of unrated orders, perhaps indicating that ordering physicians learned how to use DSS more effectively over time because more of their orders could be linked to guidelines. Among rated orders during the intervention period, between about 2 and 10 percent of initially inappropriate orders were changed or canceled across conveners, with the exception of one convener, which had an 18 percent cancellation rate. Physicians with a high ordering volume of 50 or more advanced diagnostic imaging procedures over the course of the demonstration (about 2 procedures or more per month) might be expected to have higher rates of appropriate orders relative to those who ordered fewer procedures. Yet after analyzing thousands of orders in the low ordering volume group and high ordering volume group, changes in the rate of appropriately rated orders between the baseline and intervention periods were not more frequent for physicians with a high ordering volume at most conveners, indicating that greater use of DSS does not have a discernable effect on the likelihood of appropriately ordering advanced diagnostic imaging.

Since more than 60 percent of orders were unrated, examining the trends of rated orders alone does not account for the impact of the intervention on all orders. Therefore, the evaluation modeled the probability that the typical ordering physician at each convener would order an advanced diagnostic imaging procedure that would be unrated, inappropriate, uncertain, or appropriate. For conveners with an increase in the probability of an appropriate order between baseline and intervention periods, the probability of entering an unrated order still ranged from about 30 percent to above 80 percent. For conveners with a decrease in the probability of an appropriate order between baseline and intervention, there was a corresponding increase in the probability of an unrated order. Had only rated orders been analyzed for these conveners, the percentage of appropriate orders would have increased. Thus, the substantial share of unrated orders for each convener inhibits drawing definitive conclusions about the impact of exposing physicians to appropriateness guidelines through DSS on ordering advanced diagnostic imaging.

Trends in Utilization

The evaluation examined trends in advanced diagnostic imaging utilization starting January 1, 2009—more than two years before the beginning of the demonstration—to November 30, 2013—two months after the close of the demonstration. Overall, the trends in advanced diagnostic imaging utilization did not noticeably differ for demonstration and comparison physicians before and during the demonstration, nor did they noticeably differ when stratified by convener or physician specialty type.

Propensity-weighted, difference-in-differences multivariate regression models were used to measure the physician-level effect at each convener of exposure to appropriateness guidelines through DSS during the intervention period relative to a comparison group and two separate preceding time periods—the approximately two-year pre-demonstration period and the 6-month baseline period at the start of the demonstration. In the model with the two-year pre-demonstration period, the estimated change in utilization was statistically significant for physicians within only two conveners, resulting in 1 to 2 fewer advanced diagnostic imaging procedures per 100 beneficiaries who had an office visit or any procedure at each of these conveners (or an average of 0.01 to 0.02 fewer advanced diagnostic imaging procedures per beneficiary). Only physicians within one convener had a statistically significant reduction in utilization of the same magnitude in the model with only the baseline period. Therefore, exposing ordering physicians to appropriateness guidelines for advanced diagnostic imaging over the course of two years had no effect on utilization for physicians within most conveners, and where a statistically significant effect was found, its magnitude was very small and limited to two conveners at most.

Appropriateness and Image Results

Because generating image results would have entailed a burdensome process of adjudicating results and forcing physicians to return to update their DSS orders days or weeks after an image was furnished, a direct analysis of the correlation between appropriateness of advanced diagnostic imaging orders and their results per se could not be performed. The DSS also did not capture the physician's own reason for an order in MID, which could be used to analyze whether the physician's reason for an order corresponds with the guideline triggered by DSS, if one were triggered. These data gaps are limitations of the demonstration. Instead, an analysis was undertaken of whether feedback about inappropriate orders during the first 90 days of the intervention period affected the utilization of advanced diagnostic imaging in the subsequent 90 days. It might be expected that physicians who placed a high volume of orders and had a relatively high proportion of inappropriately rated orders during their first exposure to DSS feedback would, in turn, have a reduction in utilization because they would learn not to order as many advanced diagnostic imaging procedures.

The analysis was limited to the 281 physicians with a minimum of 15 orders in the initial 90 days of the intervention period (i.e. at least 5 orders per month) and at least one order in the following 90 days. Since many orders could not be rated and only a small subset of rated orders were inappropriate, the evaluation was unable to definitively measure the impact on utilization. While the number of physicians in the analysis was relatively small, these results support the evaluation's findings that receiving feedback on inappropriate orders in the context of this demonstration did not result in reductions in advanced diagnostic imaging utilization.

Physician and Patient Satisfaction

Physician and patient satisfaction during the demonstration was an implicit part of the performance standards in each convener's participation contract with CMS. If a problem with satisfaction threatened the demonstration's ability to be conducted, then the contractor operating the demonstration responded quickly to remedy it or the convener ceased participation. No problems with physician satisfaction endangered conveners' participation contracts. Because the demonstration did not affect Medicare coverage or payment policy, beneficiaries were not notified if a physician ordered an advanced diagnostic imaging procedure while participating in MID. No known beneficiary complaints were filed in connection with MID.

The evaluation sought to understand physician and patient satisfaction with the demonstration through focus groups and short questionnaires. Convener leadership and physicians roundly liked the demonstration's intent to measure and improve the appropriateness of advanced diagnostic image ordering, but they found that MID's requirements for delivering guidelines through DSS was not an effective means to improve ordering behavior. In a supplemental questionnaire for focus group participants, more than half of physicians disagreed that the appropriateness guidelines delivered through the DSS used in the demonstration were informative or useful to their practice; were helpful in talking with patients about advanced diagnostic imaging; and allowed them to stay abreast of current best practices in advanced diagnostic imaging. Even so, generalists were more likely than specialists to have a favorable opinion of the guidelines.

Entering and changing orders in DSS added time to workflows. On average, physicians reported spending 3.9 minutes ordering an advanced diagnostic imaging procedure before the demonstration but 7.2 minutes during the demonstration. They might have been willing to spend more time ordering advanced diagnostic imaging if they thought the DSSs used in the demonstration added value to their workflows, yet physicians largely did not view them as such. The DSSs used in MID were designed to check the appropriateness of an advanced diagnostic imaging procedure that a physician planned to order. Physicians said that they would have preferred to receive guidance about different imaging procedures as they were considering placing an order, rather than deciding what to order and then consulting DSS. Spending time entering an order only to learn that it could not be linked to a guideline was especially frustrating for physicians. When an order was rated, the feedback itself simply provided a link to the guidelines, rather than providing tailored feedback to suit the context of a busy day of seeing patients. Physicians also felt frustrated from receiving DSS feedback based on guidelines that did not seem to account for all clinical aspects of the patient and sometimes conflicted with their local standards of care.

Physicians using the DSS in this demonstration perceived neither a positive nor negative effect on the quality of care their patients received. More than 80 percent of physicians believed that patients were not even aware when their orders were entered through DSS. They saw the

potential of DSS to engage patients if patients insisted on receiving an advanced diagnostic imaging procedure that was inappropriate according to the guidelines, but physicians were not confident enough in the interface and guidelines themselves to use the DSS in this demonstration in that way.

Patients who received an advanced diagnostic imaging procedure ordered through DSS were aware that the order was placed through a computer but were unaware that the ordering physician received feedback on the appropriateness of the order. In fact, they generally seemed unknowledgeable that guidelines exist for ordering advanced diagnostic imaging procedures. Patients did not perceive any delays with ordering or scheduling during the demonstration.

Lessons Learned

The demonstration was designed to provide ordering physicians with real-time feedback about the appropriateness of 12 of the most commonly ordered advanced diagnostic imaging procedures in the Medicare population. This design assumed that rigorous guidelines were available for the clinical scenarios leading to orders; that these guidelines could be programmed into the DSS in this demonstration in a user-friendly fashion; and that all physicians ordering these images would benefit from increased awareness of their appropriateness. However, convener leadership and physicians who participated in focus groups questioned whether these assumptions were valid.

A common set of national guidelines was used to rate the appropriateness of advanced diagnostic imaging orders. Because no independent consensus organization had developed appropriateness principles consistent with the statute requiring the demonstration, medical professional society guidelines were solely used as the standard to rate advanced diagnostic imaging orders for appropriateness. While professional societies might seem to be best informed to produce imaging guidelines, convener leaders pointed out that they exist to advance the interests of their members and thus have a vested interest in advising that imaging be ordered, particularly in instances where strong evidence underlying the guidelines is lacking. A limited number of advanced diagnostic imaging guidelines are supported by randomized control trials or written based on clinical outcomes; many of them are based on expert opinion. Consequently, the guidelines are subject to differences in expert opinion and may not keep pace with local evidence that can fill gaps and lags in updating national guidelines. One convener estimated that 20 to 30 percent of the guidelines used in MID were in conflict with its own local standards of care. To participate in MID, conveners had to program guidelines into their DSS that were not necessarily consonant with their local standards of care. For ordering physicians, confusion might result when orders they expected would be appropriate according to local standards of care were rated uncertain or inappropriate.

Another source of confusion—as well as frustration—for ordering physicians were situations in which no guidelines exist. More than 60 percent of orders placed throughout MID could not

be linked to a guideline, either because the ordering physician inadvertently did not enter the precise information into DSS to match to a guideline or because a guideline does not exist for a particular clinical scenario. As a result, physicians or their proxies would spend two to three minutes entering orders only to be informed that those orders were "not covered by guidelines." Physicians stated that they found DSS to be a waste of their time when it indicated that their orders could not be rated. Specialists particularly found this type of feedback unhelpful because their expertise is limited to a set of advanced diagnostic imaging procedures that they order frequently.

DSS users' frustration was compounded by the DSS interface with electronic medical records used during the demonstration, which varied in the extent to which both platforms were integrated—even across practices within the same convener. Without such integration, a patient's clinical information had to be input separately into DSS, introducing the possibility that the requisite information to link to a guideline was not entered consistently. As a requirement of the demonstration, physicians had to attest to their orders—even for appropriate orders—which meant another click in the electronic ordering process. Another limitation of the demonstration occurred whenever ordering physicians were forced to close a DSS record and re-enter patient information to create a different order in response to DSS feedback or from radiologists after placing an order. Instead of enhancing workflows, the requirements for using DSS in the demonstration often slowed workflows and eroded physicians' trust in advanced diagnostic imaging guidelines.

That many DSS orders could not be rated for appropriateness highlights the challenge of programming guidelines into an electronic user interface that can reliably trigger them. The clinical scenarios that lead a physician to consider ordering advanced diagnostic imaging represent numerous permutations of patient signs and symptoms that must be mapped to guidelines in DSS. These signs and symptoms—and their various exceptions—must first be captured in the guidelines. Assuming they are, they must be translated into computer code since the professional society guidelines used in MID were not originally written to be programmed into DSS, nor were they intended to provide real-time feedback to physicians about the appropriateness of their advanced diagnostic imaging orders. Finally, ordering physicians have to input the precise combination of clinical information into DSS to link to a guideline.

Conveners had to specially program the guidelines into the DSS used in the demonstration and implement it within a short time period of approximately nine months. The demonstration allowed each convener to procure its own DSS with the requirement that it be programmed with the common set of guidelines to MID. Conveners employed one of two main types of DSS designs. In one, users selected the patient characteristics and clinical indications for an order, which in turn were linked to a guideline variant to rate the appropriateness of the order. Another design was structured such that users clicked through a series of screens that asked questions about the indications for an order, which led to the appropriateness rating. This combination of flexibility in DSS design and rigidity in content meant that conveners could program the

guidelines differently and users could arrive at different appropriateness ratings for the same clinical scenario depending on how they entered clinical information. That the percentage of orders "not covered by guidelines" varied more than three-fold across conveners is evidence that the DSSs and the way they were used were not uniform throughout the demonstration.

Even DSS users at the same convener did not necessarily have a uniform experience entering orders and receiving appropriateness ratings. Convener leadership reported difficulty in training physicians to use non-intuitive user interfaces that were not integrated into electronic medical records. A user might fail to trigger a guideline because the interface was not nuanced enough to incorporate more-detailed clinical information or the exact clinical indication, or constellation of indications, used to map to the guideline was not selected. Users might learn that entering a certain combination of indications always produced an appropriate rating and so they simply entered what was needed to obtain an appropriate rating. Or, users might interpret the same appropriateness rating differently, leading to artificial variation in the number of changed orders.

According to convener leadership's informal feedback from physicians, the terminology used for each category in the appropriateness ratings was not plainly understandable nor provided meaningful information to ordering physicians. The appropriateness ratings were presented in a range from 1 to 9, where ratings 1 to 3 were "inappropriate;" 4 to 6 were "uncertain;" and 7 to 9 were "appropriate." The categories and their ranges reflect the way the guidelines were written, rather than based on the content and strength of the evidence supporting a linked guideline. Consider an imaging order with, for example, a rating of 6—in the "uncertain" range but close to being rated "appropriate." A physician might legitimately ask whether the order was rated "uncertain" because it might be inappropriate or appropriate depending on a patient's unique condition. Or was it "uncertain" because the evidence was ambiguous about advising one way or the other? Or did it indicate gaps in the evidence? Or was it really close to being "appropriate"? Although DSS feedback in the demonstration was linked to the guidelines triggering an appropriateness rating, users who wished to consult the guidelines themselves would usually have to search an electronic document with many pages for the guidelines of interest, rather than presenting the guidelines as a tailored summary explaining why an order was adjudicated a certain way. Few physicians in focus groups reported even consulting the guidelines.

In sum, while there are limitations of MID, it offers lessons that can be learned and suggests areas for improvement with integrating appropriateness criteria into tools designed to assist physicians with medical decision-making.

Recommendations

The statute requires the Secretary to submit to Congress a report containing the results of the demonstration evaluation, together with recommendations for such legislation and administrative action, as the Secretary determines appropriate. RAND's report makes several suggestions for addressing the challenges noted with MID. Since this demonstration was completed, the

Protecting Access to Medicare Act (P.L. 113-93) (PAMA) was enacted on April 1, 2014. Section 218(b) of such Act amended section 1834 of the Social Security Act (42 U.S.C. 1395m) by adding a new subsection (q), which established a program designed to promote the use of appropriate use criteria for applicable imaging services by ordering and furnishing professionals in applicable settings. Ordering professionals would have to consult a qualifying decision support mechanism equipped with appropriate use criteria starting in 2017. Because the PAMA provision is just beginning to be implemented, there are no recommendations for legislation or administrative action. The evaluation of MID will be taken into account as the PAMA provision is implemented.

Abbreviations

ACR	American College of Radiology
ARF	Area Resource File
BETOS	Berenson-Eggers Type of Service
CI	confidence interval
CMS	Centers for Medicare and Medicaid Services
CPOE	computerized physician order entry
CPT	current procedural terminology
CT	computed tomography
CTA	computed tomography angiography
DRA	Deficit Reduction Act
DSS	decision support system
EMID	Medicare Imaging Demonstration evaluation
HER	electronic health record
HHA	home health agency
FFS	fee-for-service
HCC	Hierarchical Condition Category
HCPCS	Healthcare Common Procedure Coding System
HHS	Department of Health and Human Services
MedPAC	Medicare Payment Advisory Commission
MID	Medicare Imaging Demonstration
MRI	magnetic resonance imaging
NPI	National Provider Identifier
NPPES	National Plan and Provider Enumeration System
ROE	radiology order entry
SNF	skilled nursing facility
SPECT-MPI	single photon emission computed tomography myocardial perfusion imaging
TIN	Tax Identification Number

This page is intentionally blank.

Section I: Report Overview

This section includes Chapter 1, which provides a background for and description of the demonstration and its evaluation.

This page is intentionally blank.

1. Background and Description of the Demonstration and Its Evaluation

Increasing use of advanced medical imaging is often cited as a key driver of medical spending growth, including a fourfold increase in Medicare program spending between 1995 and 2005. Although recent trends show a decline in utilization rates for advanced imaging, several studies have found that many imaging requests are inappropriate, risking harm to patients through radiation exposure and increasing health care costs. Inappropriate medical imaging also imposes other, less commonly considered risks, including downstream tests and procedures resulting from misdiagnosis or from findings that are otherwise benign. These downstream effects have little value at best, and can be harmful at worst.

Several approaches have been implemented or recommended to curb the use of inappropriate imaging procedures. One approach uses computer tools known as decision support systems (DSSs) that provide physicians and patients with feedback on the appropriateness of treatment options being considered during patient encounters. DSS tools synthesize existing evidence on the effectiveness of each procedure, including appropriateness criteria developed by physician specialty societies, so physicians can base decisions on the current evidence at the point of care. Despite evidence of potential effectiveness, current adoption and use of DSSs are limited.

Beginning in January 2011, the Medicare Imaging Demonstration (MID) tested a specific type of DSS on a large scale to provide insight into the challenges and opportunities for deploying DSSs in Medicare fee-for-service (FFS) settings, an area identified with high rates of inappropriate use. The design was expected to permit examination of appropriateness of advanced imaging orders across geographic areas, specialties, and practice settings for Medicare FFS patients. The Centers for Medicare and Medicaid Services (CMS) contracted with five "conveners" that aggregated participating practices and served as the point of contact for implementation of the demonstration. Twelve advanced imaging procedures were selected for the demonstration on the basis of (1) a high volume of orders; (2) availability of clinical appropriateness criteria developed or endorsed by a medical specialty society; and (3) variation in utilization rates between urban and rural geographic areas and states.

The overall goal of this project is to evaluate the MID to assess rates of physician ordering of advanced diagnostic imaging services, whether the DSS tool is a viable option for exposing physicians to professional guidelines when ordering advanced imaging, and how, if at all, DSS exposure is associated with changes in the use of advanced imaging. Based on these analyses, we also provide recommendations for the consideration of policymakers.

In the remainder of this introduction, we provide a review of the literature on advanced imaging utilization; then describe the MID in more detail, including the general approach and requirements for the demonstration; and finally, we describe the structure of this report.

1.1 Literature Review

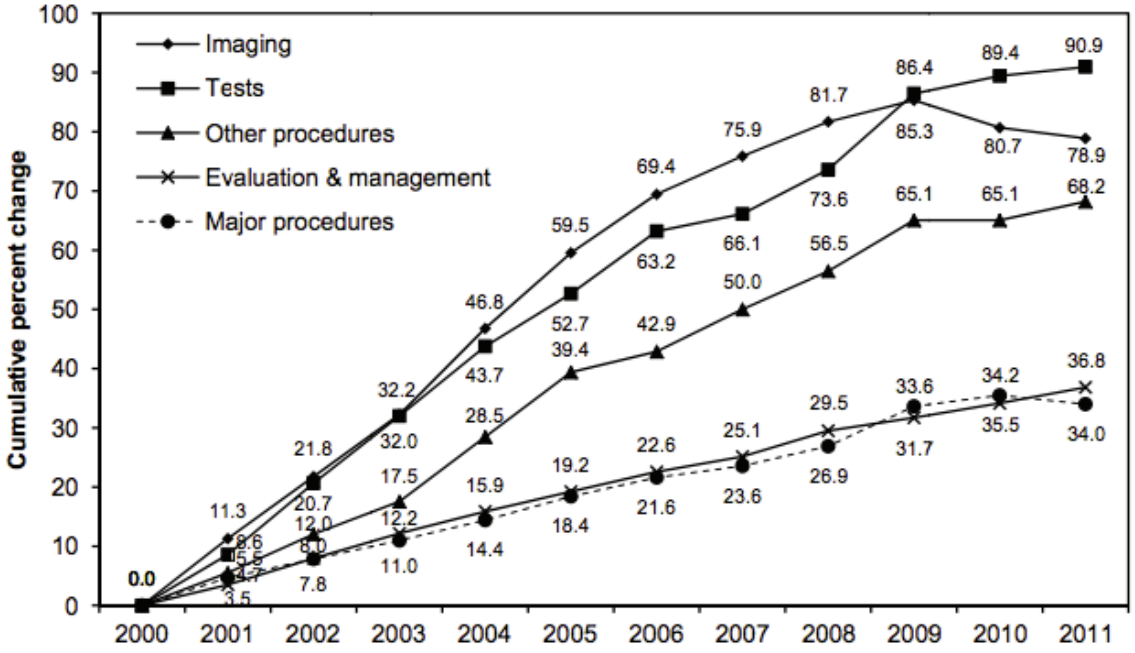
1.1.A. Growth and Decline in Utilization

Sizable growth in utilization of advanced imaging during the early 2000s elevated image volumes to historic highs. From 2000 to 2006, computed tomography (CT), magnetic resonance imaging (MRI), and nuclear medicine utilization for Medicare FFS beneficiaries grew across all major places of service. Private offices reported compound annual growth rates of 17.5 percent, 14.9 percent, and 18.4 percent, respectively, for CT, MRI, and nuclear medicine. Hospital outpatient departments saw growth as well, albeit with less intensity. During this period, the volume of CT, MRI, and nuclear medicine procedures conducted in hospital outpatient settings increased by 8.6 percent, 11.3 percent, and 4.2 percent, respectively (Levin, Rao, and Parker, 2012).

Recent evidence suggests that the growth in utilization of advanced imaging decelerated around 2006. Where advanced imaging annual growth was 13.7 percent from 2000 to 2006, there was an annual decline of 4.5 percent in spending for advanced imaging from 2006 to 2011 (Lee, Duszak, and Hughes, 2013). From 2006 to 2009, annual growth rates of CT and MRI utilization by Medicare beneficiaries declined from 7.1 percent to 1.4 percent and from 14 percent to 2.6 percent, respectively (Lee and Levy, 2012). Decreased growth also occurred in the commercial population, as annual growth rates of CT use dropped to 3.1 percent from 2006 to 2009. One study found a reduction in the per capita utilization growth rate for Medicare beneficiaries from 2004–2006 to 2006–2008 consistent with a 2.0-percentage point deceleration in the use of office and a 0.7-percentage point deceleration in the use of outpatient hospital musculoskeletal imaging per 1,000 Medicare beneficiaries (Kennedy and Foreman, 2012).

Positive growth in advanced imaging stopped in 2010 (Figure 1.1), with a drop in the CT rate from 637 to 626 per 1,000 Medicare beneficiaries, equaling a –1.7 percent change in CT utilization (Levin, Rao, and Parker, 2012). While private-office settings exhibited the largest negative change (–7.8 percent), hospital outpatient departments reduced CT utilization by –3.6 percent (Levin, Rao, and Parker, 2012).

Figure 1.1. Growth in Volume of Physician Fee Schedule Services per Beneficiary, 2000–2011



SOURCE: Medicare Payment Advisory Commission (MedPAC), 2013.

Lang et al. (2013) examined both MRI and CT utilization rates from 2000 to 2009 using the Medical Expenditure Panel Survey data. Combined MRI and CT rates increased during that period from 64.3 to 109.1 per 1,000 person years. However, the authors note that utilization annual growth rate was near 0 percent for the latter half of their study, in contrast to the 15-percent annual growth observed during the first five years. Medicare beneficiaries' utilization rate ranged from 227 to 258 per 1,000 person years and was the highest utilization rate compared to utilization rates for patients with other types of insurance during the study period.

1.1.B. Reasons for Decline in Utilization

The slowdown in volume of advanced imaging coincided with the onset of multiple policies. One of the most important factors was the Deficit Reduction Act (DRA) of 2005, which reduced payments for office-based imaging to hospital outpatient levels. One study found a correlation between lower reimbursement for office-based imaging procedures, as outlined by the DRA, and imaging growth rates in office settings for Medicare beneficiaries. By analyzing pre-DRA (2004 to 2006) and post-DRA (2006 to 2008) data, the authors found that each 1-percent cut in reimbursements for radiologists and nonradiologists resulted in utilization growth slowing by 0.2 percent (Kennedy and Foreman, 2012). In addition, payment cuts in the DRA may have led to larger declines in utilization in imaging centers and physician offices than in hospital outpatient settings (Lee and Levy, 2012).

Other policies—such as prior authorization, increased patient cost-sharing, and the use of radiology DSSs—may be more focused on the privately insured population but may have spillover effects in reducing advanced imaging utilization for the Medicare population. However, empirical evidence for spillover effects for the Medicare population have not been systematically documented.

1.1.C. Applying Appropriateness Methodology to Advanced Diagnostic Imaging

Randomized clinical trials are generally agreed to be the highest quality evidence for generating data to support the effectiveness of interventions such as the use of advanced imaging. However, trial evidence is available for only a fraction of the interventions that are employed in clinical practice. In 1980, RAND introduced the concept of appropriateness ratings as one method for systematically synthesizing data from evidence reports and expert opinion where randomized trials were inadequate or unavailable (Fink et al., 1984; Brook et al., 1986). The method emphasized the importance of selecting a diverse set of experts who are free of conflicts of interest, providing a high-quality literature review using state of the art literature synthesis techniques, developing a precise set of specification of reasons for orders, and assigning ratings according to expected outcomes for a cohort of patients (not for a particular individual).

With the RAND Appropriateness Methodology (Brook et al., 1986), indication lists are generated based upon procedure-specific medical literature. Indications are designed to be precise enough so that experts assigning appropriateness ratings would consider the benefits and risks of procedure use for patients who are homogeneous with respect to the reason for the order. The number of indications for a procedure is determined by the number of groupings that are distinct with respect to risks and benefits. Procedures are precisely defined (for example, by current procedural terminology [CPT] codes) and specify variations, such as with, without, or with and without the use of contrast. Patient characteristics are categorized according to potential reasons for the procedure, including well- specified symptoms (for example, describing the type, duration, intensity, and precipitating and relieving factors), past medical history (for example, describing prior diagnoses and associated timing), and signs and diagnostic studies (including the persistence, severity, and scope of the abnormalities).

Experts are provided evidence-based literature reviews along with rating sheets with instructions for using the literature and for assigning appropriateness ratings based upon their best clinical judgment (informed by experience and literature) while considering an average cohort of patients with the problem described by the indication. Ratings were considered "appropriate" when the expected health benefit (i.e., increased life expectancy, pain relief, reduction in anxiety, improved functional capacity) exceeded the expected negative consequences (i.e., mortality, morbidity, anxiety of anticipating the procedure, pain produced by the procedure, time lost from work) by a sufficiently wide margin that the procedure was worth doing. Ratings were considered "inappropriate" when negative consequences were likely to

outweigh benefits if the procedure were performed (compared with not being performed). Both appropriate and inappropriate ratings were assigned according to the expected outcome for a cohort of patients who would be similar with respect to the reason for the order. For example, while considering 1,000 patients with low back pain and no signs of dangerous pathology, the expected outcome would be considered for patients receiving the procedure in question compared those with not receiving it. An appropriate rating would be assigned if the expected benefit of performing the procedure outweighed the risk, and an inappropriate rating would be assigned if the expected risk outweighed the benefit. With the original RAND appropriateness ratings, costs were not included in the assignment of appropriateness (Park et al., 1986; Kahn et al., 1988a, Kahn et al., 1988b, Winslow, 1988; Park et al., 1989). Further definitions and details of appropriateness ratings are described in Section 8.2.D.)

Building on RAND's appropriateness methodologies, specialty societies have developed appropriateness criteria using a structured process to integrate available evidence from clinical studies with clinical judgment provided by an expert panel (Fitch et al., 2001). Several studies have found that many imaging studies are inappropriate, particularly among older patients (Hendel 2009, 2010). In part, this may reflect lack of evidence among older patients, as well as lack of evidence for comparative effectiveness of imaging strategies.

Inappropriate use of imaging can harm patients and increase health care costs. The clinical risks with advanced imaging are significant, with concerns about radiation exposure gaining traction in debates among physicians (Brenner and Hall, 2007; Fazel et al., 2009). Inappropriate medical imaging also imposes other, less commonly considered risks, including downstream tests and procedures that have little value or are harmful and that result from misdiagnosis or from findings that are otherwise benign (Fisher, 2003).

1.1.C. Prevalence and Consequences of Inappropriate Utilization of Advanced Diagnostic Imaging

Despite recent trends showing a slowdown in the rate of advanced image prescribing, there is widespread agreement that a sizable proportion of advanced imaging studies remain inappropriate. The development and dissemination of appropriateness criteria by professional societies, such as the American College of Radiology (ACR) and the American College of Cardiology, have provided tools to help characterize the rates of inappropriate use of advanced imaging, but most published studies assess the prevalence of inappropriate use only within a single institution (Gibbons et al., 2008; Miller et al., 2010; Lehnert and Bree, 2010), which means that these findings may not be generalizable to other practice settings. The majority of appropriateness studies involve cardiology, where rates of inappropriate use of nuclear medicine procedures range consistently between 13 percent and 18 percent (Gibbons et al., 2008; Hendel et al., 2010; Mehta et al., 2008). Inappropriate use of coronary CT angiography has been estimated to be nearly 17 percent (Miller et al., 2010). One study that assessed inappropriate use of CTs and MRIs overall found rates as high as 26 percent.

Inappropriate use of advanced imaging has drawn the attention of policymakers not only because of its escalating costs, but also because of its potential risks to patients' health. Certain types of advanced imaging expose patients to high levels of ionizing radiation, and cumulative exposure significantly increases a patient's lifetime risk of developing cancer (Brenner and Hall, 2007). Patients may also experience reactions to contrast media, ranging from relatively minor side effects, such as pruritus, to life-threatening renal failure (Maddox, 2002). Other potential harm from imaging includes patient anxiety or a cascade of subsequent diagnostic tests and interventions. Medical care provided on the basis of incidental findings of an advanced image may have limited benefit, lower patients' quality of life, or be harmful (Deyo, 2002). Studies have shown that awareness of the potential harms from advanced imaging varies by specialty and that physicians do not always discuss these risks with patients (Lee et al., 2004).

Estimated rates of inappropriate utilization may not be entirely accurate, however, because in many cases, published appropriateness criteria are not applicable to all types of patients seen in routine practice, and in many cases insufficient evidence exists for experts to achieve consensus on the appropriateness of a particular indication. In one study, appropriateness ratings for nearly 29 percent of patients with indications for cardiac CT procedures could not be determined (Murphy et al., 2010), while in another, up to 46 percent of patients with indications for coronary CT angiography could not be assigned appropriateness ratings (Miller et al., 2010). Data compiled at a preauthorization center for MRI requests indicate that, as of 2002, ACR appropriateness criteria could not be applied to 41 percent of MRI imaging requests (Levy et al., 2006).

By contrast, all but 4–10 percent of patients could be mapped to appropriateness criteria for single photon emission computed tomography myocardial perfusion imaging (SPECT-MPI), suggesting that for some advanced imaging procedures, these latter criteria are broadly applicable to patients typically found in most practices (Gibbons 2008; Mehta et al., 2008). Mehta's (2008) analysis of reasons for SPECT-MPI orders for 1,209 patients in a single institution were examined overall and by specialty. While no significant difference was noted across images referred by cardiology, internal medicine, or emergency medicine practices, a significantly higher rate of inappropriate orders was associated with referral for anesthesia pre-operative evaluations ($p<0.002$). For example, between 7 percent and 11 percent of patients being considered for SPECT-MPI may have indications that generate "uncertain" ratings (Gibbons et al., 2008; Mehta et al., 2008; McCully et al., 2009), and rates for cardiac computed tomography angiography (CTA) may be similar (Gibbons et al., 2008). Approximately 15 percent of patients in one study who were evaluated for cardiac CT had uncertain indications (Murphy et al., 2010).

Gaps in appropriateness criteria often prompt local providers to augment the criteria produced by professional societies with their own decisions on appropriateness. Studies have shown that clinicians use appropriateness criteria far less often than other resources, such as

specialist consults and UpToDate (Wolters Kluwer Health, undated), to guide the management of their patients (Bautista et al., 2009).

1.1.D. Interventions Designed to Limit Inappropriate Use of Advanced Diagnostic Imaging

A number of strategies have been used to limit the inappropriate use of advanced imaging, which may account for the declines in volume of advanced imaging procedures in the latter half of the 2000s. These strategies fall into five main categories: payment policy, accreditation, utilization management, public awareness and education campaigns, and decision support. We consider each of these in turn, and then focus on previous studies of DSSs in the final section.

1.1.D.1. Payment Policy

The DRA of 2005 sought to eliminate the payment differentials and other types of mispricing that have provided perverse incentives encouraging clinicians to order advanced imaging. One provision effectively equalized payments for the technical component of imaging fees to reduce the incentive to provide imaging in a physicians' office vis-à-vis other outpatient settings.[2] The DRA also reduced technical component payments by 25 percent for images ordered for contiguous body parts on the same day (which was subsequently reduced to 50 percent by the Affordable Care Act).[3] Also in 2007, CMS updated estimates of office expenses, including the costs required to acquire and maintain imaging equipment, which affected the calculation of the technical component of the physician fee schedule. This change led to increased payments for evaluation and management services at the expense of imaging services.

In the first year of DRA, Medicare spending on imaging decreased by 12.7 percent (Hillman and Goldsmith, 2010a). Some data suggest that utilization of CT, MRI, and nuclear medicine declined in 2007—the first full year following the enactment of DRA—from a growth rate of 10.2 percent over the six-year period prior to 2007 to 1.7 percent during 2007 (Levin, Rao, and Parker, 2010). Between 2008 and 2009, growth in diagnostic imaging averaged only 2 percent, compared to an average growth of 6.3 percent between 2004 and 2009 (MedPAC, 2011). The revaluing of technical fees led to dramatic changes in the organization of cardiology practices, which relied heavily on imaging as mentioned previously. The CEO of the American College of Cardiology estimated in 2010 that these payment changes may have reduced the share of cardiologists working in private practice by one-half in just one year (Harris, 2000).

---

[2] The technical component covers the costs of the nonphysician clinical staff who perform the test, as well as medical equipment, medical supplies, and overhead expenses.

[3] For example, for a CT scan involving the chest, abdomen, and pelvis, the first image would be reimbursed at a full rate but the second and third ordered body part images would be reimbursed at 75 percent of the scheduled payment. (Hillman and Goldsmith, 2010a).

1.1.D.2. Accreditation

Accreditation programs for practice facilities that render advanced imaging have the goal of ensuring that providers who perform imaging procedures have appropriate qualifications, use high-performing equipment, and implement robust quality-control processes. (Bernardy et al., 2009). Beginning January 1, 2012, CMS required all providers of advanced imaging to seek accreditation from one of three national accreditation organizations, including the ACR, the Joint Commission, or the Intersocietal Accreditation Commission.

1.1.D.3. Utilization Management Policies

Other policies focus on reducing inappropriate utilization directly. Utilization management typically includes prior authorization and prior notification policies. *Prior authorization* policies block imaging orders until approval is granted, whereas *prior notification* policies require providers to submit information on indications at the time of ordering but stop short of imposing barriers to ordering. Prior-authorization policies can substantially reduce utilization of advanced imaging (Blachar et al., 2006), presumably by reducing the rates of inappropriate orders relative to appropriate orders—but to date, no rigorous studies have evaluated the impact of prior authorization programs on the appropriateness of advanced imaging (MedPAC, 2011). One factor that might affect the success of these policies is that physicians begin to learn which diagnoses lead to approval (Mitchell and Lagalia, 2009). Moreover, results from existing studies indicate that, despite significant decreases initially, the impact of these policies may weaken over time.

1.1.D.4. Feedback Reports on Radiology Ordering

Many health care organizations use audit and feedback reporting as a cornerstone of their quality improvement initiatives. CMS has also embraced feedback reporting on a large scale as it begins to disseminate reports on quality and cost measures that include benchmarks to clinicians participating in the Medicare program (U.S. Congress, 2010). While a rich literature exists on the effectiveness of audit and feedback reporting in changing clinical practice (Jamtvedt et al., 2003; Grol and Grimshaw, 2003), the evidence on the impact of feedback reporting on radiology ordering behavior, specifically, is less clear. In particular, federally sponsored public reporting and feedback initiatives do not typically include a large number of radiology measures. Moreover, the impact of participating in radiology-focused quality improvement initiatives that allow benchmarking to peers, such as the American College of Radiology's National Radiology Data Registry, has not been studied extensively to date (ACR, 2014).

1.1.D.5. Radiology Order Entry

Computerized radiology order entry (ROE) entails an electronic method of radiology ordering that requires the ordering provider to specify the requested image type along with clinical data on symptoms or indications. ROE may or may not include additional decision

support, such as feedback on the appropriateness of the order based on the specified modality and indications. Introduction of computerized ROE without decision support may influence advanced imaging ordering by promoting reflection on indications for each order at the time of ordering. In one study using a pre-post design, the introduction of ROE without DSSs led to a reduction in the growth of ultrasound imaging from 9 percent to 4 percent per year (Sistrom et al., 2009).

A recent systematic review included 14 studies that assessed the impact of ROE systems (Georgiou et al., 2011). Similar to the literature on the effects of prior authorization, the majority of these studies used relatively weak designs, including pre-post studies without parallel control groups. Nine studies included some form of decision support, while the remaining studies did not explicitly report decision support features. Because six of the 14 studies took place in critical care settings, findings from this collection of studies may not be applicable to the MID. In general, the authors of the review concluded that the body of evidence supporting the impact of ROE is small.

1.1.D.6. Public Awareness and Educational Campaigns

Other factors that could explain the recent drop in imaging utilization include educational campaigns and public awareness. For example, Hillman and Goldsmith (2010b) recommend inclusion in the medical school curriculum of specific training in when to consider imaging, what is appropriate imaging, and how to consult with a radiologist as a supplement to how to interpret images. Increased public focus on the potential risks of unnecessary imaging reported by news media and peer-review publications coincided with the reduction in imaging rates (Lee and Levy, 2012; Levin, Rao, and Parker, 2012). Educating providers and patients about appropriate use of imaging is a recent collective effort by specialist groups. The *Imaging Wisely* and *Choosing Wisely* campaigns are two leading examples of industry-led efforts to encourage appropriate use of care and advanced diagnostic imaging services.

Imaging Wisely is a program that began in 2009 "with the objective of lowering the amount of radiation used in medically necessary imaging studies and eliminating unnecessary procedures" (ACR, 2010). The effort is supported by the major radiology specialty groups, including the ACR, and partners with imaging equipment vendors to promote education regarding appropriate radiation dose levels (Brink and Amis, 2010).

Choosing Wisely is led by the American Board of Internal Medicine Foundation and encourages physicians and patients to engage in conversations to achieve care that is "supported by evidence; not duplicative of other tests or procedures already received; free from harm; and truly necessary" (ABIM Foundation, 2012a). The initiative has more than 60 partners, including medical specialty associations, *Consumer Reports,* and the American Association for Retired Persons (AARP). Partnering specialty societies generate a "top five" list of medical services whose utility should be questioned by both the patient and provider (Gliwa and Pearson, 2014). As of August 2013, more than 135 services have been listed. More than 50 specialty societies

have each produced lists that reflect evidence-based recommendations on behalf of the specialty society. The ACR, a member of the *Choosing Wisely* campaign, recommends the following:

- Don't perform imaging for an uncomplicated headache.
- Don't image for suspected pulmonary embolism without moderate or high pretest probability.
- Avoid admission or preoperative chest X-rays for ambulatory patients with unremarkable history and physical exam.
- Don't perform CTs for the evaluation of suspected appendicitis in children until after ultrasound has been considered as an option.
- Don't recommend follow-up imaging for clinically inconsequential adnexal cysts.

Other radiology subspecialty societies (e.g., Society of Nuclear Medicine and Molecular Imaging, Society for Cardiovascular Magnetic Resonance, American Society of Nuclear Cardiology) have also contributed lists to the *Choosing Wisely* campaign (ABIM Foundation, 2012b). Initiatives such as *Imaging Wisely* and *Choosing Wisely* represent important current mechanisms that may be contributing to changing trends in advanced diagnostic imaging by causing clinicians to carefully assess the evidence supporting the use of advanced imaging procedures.

1.1.E. Effectiveness of DSSs in Promoting Appropriate Use of Advanced Imaging in Outpatient Settings

We identified a small number of studies that sought to evaluate the impact of introducing DSSs on the appropriateness of ordering behavior in outpatient settings (Sistrom et al., 2009; Rosenthal et al., 2006; Vartanians et al., 2010; Blackmore, Mecklenburg, and Kaplan, 2011; Solberg et al., 2010; Bowen et al., 2011; Curry and Reed, 2011; Ip et al., 2012). Two of the most widely cited studies involved an evaluation of the implementation of ROE with decision support at the Massachusetts General Hospital (Sistrom et al., 2009, Rosenthal et al., 2006). For this intervention, feedback to physicians included "utility scores" on a nine-point scale. Physicians who disregarded the DSS's recommendation were required to document their reasons, and physicians having a large number of low-utility orders were required to participate in educational sessions. The authors of this study note that hospital staff supplemented the ACR appropriateness criteria with locally developed criteria where they were lacking, using an expert panel that generated appropriateness ratings by consensus.

Over a three-year period following the introduction of a DSS at the hospital, rates of CT orders decreased from 12 percent to 1 percent annually, and utilization of MRIs decreased from 12 percent to 7 percent per year. Low-utility orders decreased from 6 percent to 2 percent overall, and the reductions were greatest for primary care physicians. More than 90 percent of low-utility orders were followed through to completion. In nearly 50 percent of cases, physicians documented that their reason for continuing with the order (a requirement for completing the order) was that the order was recommended by a specialist; in nearly 25 percent of cases,

physicians indicated disagreement with the guidelines. In a separate study, Vartanians et al. (2010) assessed the impact of a policy that barred nonclinician staff from completing orders that were initially given a low utility rating by the DSS and found that low-utility orders decreased from 5.4 percent to 1.9 percent.

Using a retrospective cohort design, Blackmore, Mecklenburg, and Kaplan (2011) demonstrated the effectiveness of DSSs following the introduction of one within Virginia Mason Medical Center. The authors showed that the DSS was associated with a decrease in the utilization of lumbar MRIs for low back pain (risk ratio, 0.77; 95-percent confidence interval [CI] 0.87-0.67; p=0.0001), head MRIs for headache (risk ratio, 0.76; 95-percent CI 0.91-0.64; p=0.001), and sinus CTs for sinusitis (risk ratio, 0. 73; 95-percent CI 0.82-0.65; p<0.0001). The authors attributed the success of the intervention to its targeted nature (focusing on three imaging modalities for three specific indications), internal development and vetting of the guidelines, and automatic rejection of orders that did not conform to approved indications. The intervention took place in an integrated delivery system and was accompanied by educational sessions, conferences, and personal outreach—factors that might have accounted for the program's success.

A third DSS demonstration occurred within the HealthPartners Medical Group, a multispecialty medical group in Minnesota (Solberg et al., 2010). This demonstration resembles MID by allowing physicians to override recommendations of the DSS and by excluding other educational programs from the demonstration. In addition, no financial incentives were linked to a physician's ordering patterns. In this study, introduction of a DSS led to reductions in use for two of the three highest volume orders: head CTs (37 percent) and spine MRIs (20 percent), but rates of head MRIs increased slightly. Improvements in the appropriateness of orders were observed for head MRIs only. One major limitation of this study was the short duration of follow-up, which was limited to a two-month period.

A fourth study by Ip et al. (2012) examined whether an integrated imaging computerized physician order entry (CPOE) system with embedded decision support that was developed between 2000 and 2010 could be accepted by clinicians in their routine clinical practice. An important goal of the system from its conception was a feasible integration into the information technology framework of the health care system. Iterative informal feedback was collected from users and integrated into the CPOE system to optimize workflow. Defining their primary outcome measure as the proportion of all imaging examinations performed with orders electronically created by authorized system-level providers, and the proportion of all imaging examinations performed electronically signed by authorized providers even when orders were first entered by staff proxies, the goal was at least 90-percent adherence to these measures. Examination of a total of 4.1 million diagnostic imaging studies performed following request by 14,899 unique users showed that target goals were met. The most frequently cited helpful workflow features included the integration of an online scheduling module, the ability to place electronic signatures, an intuitive user interface, implementation of user-specified macros,

elimination of redundant data entry, integration of the CPOE with third-party payer's preauthorization processes where possible, and access to real-time decision support.

While these four studies suggest some benefit of DSSs in reducing inappropriate utilization of advanced imaging, the generalizability of their findings is uncertain. Two studies involved unique types of provider organizations, and two involved implementation in an academic medical center. These settings may not be representative of a typical outpatient clinical practice. In each of these studies, the guidelines used were extensively vetted at each institution, and two of the implementations were accompanied by educational outreach. None of the studies used a prospective comparison group. Finally, studies of DSSs that yield positive outcomes may be more likely to be published than studies indicating no impact, potentially providing a distorted picture of the overall effectiveness of DSSs.

Not all effects of DSSs have been positive. One evaluation of DSSs conducted at a Children's Hospital in Winnipeg, Canada, found that the system was not well accepted by physicians (Bowen et al., 2011). Only 19 percent of orders had relevant guidelines. Eleven percent of these were rated as inappropriate according to the guidelines. Among these inappropriately rated orders that were associated with an alternative recommendation (for either an alternative or no image), physicians followed the DSS guidance only 2 percent of the time. Physicians felt the DSS was "too generic" and not applicable to the complex or high-risk patients they treated. The timing of feedback was also cited as a barrier because physicians typically had already discussed their recommendations with the parents of their patients by the time they entered data into the DSS. In this demonstration, physicians acknowledged "cheating" by avoiding input of information that would have resulted in additional prompts, thus interrupting their workflow.

In another evaluation of a DSS implemented in a rural community family practice clinic, a higher but still limited proportion of orders (58 percent) were addressed by the Canadian Association of Radiologists guidelines (Curry and Reed, 2011). A total of 24 percent of orders were determined to be inappropriate or unnecessary, and physicians followed DSS suggestions on 25 percent of these orders, although there was significant between-provider variation in willingness to change orders. The extra time required to enter data was a source of complaints, although the percentage of physicians who described the DSS as disruptive declined from 40 percent to 16 percent over time. Half of physicians participating in this study questioned the validity of the guidelines.

The latter study brings up a critically important gap in the literature involving the use of appropriateness criteria within a DSS. As early as 2008, the ACR documented their more than ten-year history developing Appropriateness Criteria® (Jamal and Gunderman, 2008). Furthermore, they subjected the process to an internal evaluation by fourth-year medical students endowed with recent exposure to a wide variety of clinical disciplines, extensive training about concurrent and planned medical practice innovations, and emerging information technologies. While the students endorsed the Colleges' program to develop appropriateness criteria, they also

noted a number of opportunities for improvement. Recommendations included (1) specifying non-radiologic alternatives as a supplement to radiologic alternatives for inappropriately ordered images; (2) adding cost, availability, and adverse effects potentially associated with recommended images; (3) incorporating information about the sequencing and timing of imaging examinations; (4) further study of the best ways to present appropriateness ratings; and (5) inclusion of a broader set of clinician users into the development of the criteria.

Few studies have validated the advanced imaging appropriateness criteria that are widely used in practice. Solberg et al. (2010) examined the impact of DSSs on actual patient outcomes. That study found that improvements in the appropriateness of MRIs were accompanied by increases in the rate of "positive findings" for only one of the two MRI procedures (Solberg et al., 2010). The authors concluded that the validity of the ACR guidelines is unclear. This study found no difference between orders rated as appropriate or otherwise with respect to the proportion of images with abnormal results or on the likely impact on patients. Nevertheless, substantial evidence in fields other than radiology have documented that better processes of care are associated with improved patient outcomes (Kahn et al.,1990; Higashi et al., 2005, Kahn et al., 2007) . As more studies of the effectiveness of DSSs for advanced imaging emerge, we can expect to learn more about how, if at all, patient outcomes change as a function of use of guideline-supported images.

The literature review indicates that a better understanding is needed of the effectiveness of using DSS-provided guidelines to reduce the frequency of inappropriate advanced imaging orders. In particular, it is important to understand the impact of exposing clinicians to medical specialty guidelines on ordering advanced imaging procedures *at the time of their order*. To date, only a few studies have sought to evaluate the impact of introducing DSSs on the appropriateness of ordering behavior in outpatient settings, and the findings from the few existing studies may not be generalizable across organizations and settings. The evaluation of the MID seeks to address this gap by evaluating the impact of exposing clinicians to medical specialty guidelines on the appropriateness of advanced imaging orders at the time of order. The study examines these effects across different geographic locations, specialties, and types of practices.

1.2. The Medicare Imaging Demonstration

The MID was authorized by Section 135(b) of the Medicare Improvements for Patients and Providers act of 2008. The statute instructs the Department of Health and Human Services to collect data to determine the appropriateness of services in relation to established criteria and to provide participating physicians with feedback reports that permit comparison against physician peers on utilization and adherence to appropriateness criteria. The statute emphasizes the use of DSSs to achieve these goals and prohibits the use of prior-authorization requirements. The

design was expected to permit examination of appropriateness of advanced imaging orders across geographic areas, specialties, and practice settings for Medicare FFS patients.

CMS contracted with "conveners" that aggregated participating practices and served as the point of contact for implementation of the demonstration. Many types of entities were permitted to serve as conveners. The following criteria were applied in selection of demonstration conveners: diverse mix of physician practice sizes, specialties, geographic locations, patient demographic characteristics, and practice types; minimum of five annual orders across MID's targeted advanced imaging procedures; availability of DSSs; and capacity to meet other demonstration requirements, such as data submission to the implementation contractor.

Five conveners were selected to participate in the demonstration (Table 1.1).

Table 1.1. Description of MID Conveners

Convener Name	Key Characteristics	Number of Practices	Number of Practice Sites
Brigham and Women's Hospital	Coalition of three academic systems—Brigham and Women's Hospital, Weill Cornell Medical College, and University of Pennsylvania Health System—and one integrated delivery system, Geisinger Health System	4	59
Henry Ford Health System	Comprehensive, integrated, nonprofit, managed care, health care delivery system located in Southeast Michigan	1	36
Maine Medical Center	Association of several types of practices of varying sizes including academic hospital affiliated, nonacademic hospital affiliated, and independent.	20	84
National Imaging Associates	Independent practices that have collaborated with National Imaging Associates for radiology benefit management services and have technology infrastructure to support DSSs.	8	125
University of Wisconsin Medical Foundation	Academic medical center partnered with two group practices (Meriter and Group Health Cooperative of Southern Wisconsin).	3	59

SOURCE: Key characteristics: demonstration applications. Practice and practice counts: convener workbook data (submission 00).

The number of practices per convener ranged from one to 20 at the start of the demonstration, reflecting heterogeneity in what constitutes a "practice." The participating practices are located in eight states: Maine, Massachusetts, Michigan, New York, New Jersey, Pennsylvania, Texas, and Wisconsin. All conveners list multiple sites for most practices, with the number of practice sites ranging from 36 to 125 per convener. Accordingly, conveners varied substantially with respect to the number of patients, providers, and image orders.

To optimize homogeneity in size across our analytic units, we reconfigured the five conveners contracted to participate in MID into seven subconveners; we did this by grouping convener practices that share similar MID implementation procedures. Each subconvener had its

own leadership for purposes of interacting with MID's implementation and the evaluation contractors. While subconveners shared some DSS software features with their "main" convener, each subconvener implemented MID in a manner that reflected its own workflow. Subconveners varied with respect to urbanicity, size (number of patients, providers, and practices), prior exposure to radiology order entry, decision support, and use of electronic health records.

Throughout the remainder of this report, conveners and subconveners will be referred to interchangeably as *conveners*. Thus, this report presents analyses pertinent to seven unique "conveners" comprised of a combination of the five original conveners and subconveners reconfigured to assure more comparability of ordering volumes across conveners. Descriptions of the original five conveners are shown in Table 1.1.

The MID is focused on 12 advanced imaging procedures in three advanced imaging modalities: MRIs, CTs and nuclear medicine (Table 1.2). Each of the 12 procedures includes a "family" of Healthcare Common Procedure Coding System (HCPCS) codes, which include the procedure performed with contrast, without contrast, and without followed by with contrast (with the exception of SPECT-MPI, where the family includes HCPCS codes for single or multiple studies at rest and/or stress). These 12 procedures were selected according to three criteria: (1) high volume; (2) availability of clinical appropriateness criteria developed or endorsed by a medical specialty society; and (3) variation in utilization rates between urban and rural geographic areas and states. These criteria were selected by The Lewin Group via analysis of Medicare claims data and stakeholder recommendations (medical societies and radiology benefit managers).

The intervention in the MID is designed to support two main activities: (1) exposure to medical specialty guidelines on the appropriateness of ordering advanced imaging procedures at the time of order through DSSs; and (2) receipt of periodic feedback reports to providers on their ordering patterns in comparison to their peers.

DSSs provide clinicians with the opportunity to: (1) use ROE to order advanced imaging electronically while documenting individual patients' characteristics; (2) benefit from available synthesized evidence about the images' effectiveness at the point of care; and (3) receive real-time feedback about the appropriateness of their orders and suggestions for potential changes or additions to the order.

Table 1.2. Demonstration Imaging Procedures

Demonstration Advanced Imaging Procedure	HCPCS Codes	Source of Relevant Clinical Guidelines
CT Abdomen	74150, 74160, 74170	ACR
CT Pelvis	72192, 72193, 72194	ACR
CT Abdomen and Pelvis	74176, 74177, 74178*	ACR
CT Brain	70450, 70460, 70470	ACR; American Academy of Neurology; American Academy of Otolaryngology; U.S. Headache Consortium
CT Lumbar Spine	72131, 72132, 72133	American Academy of Neurology; American College of Physicians/American Pain Society; ACR; North American Spine Society
CT Sinus	70486, 70487, 70488	ACR; American Academy of Otolaryngology
CT Thorax	71250, 71260, 71270	ACR; American Academy of Family Physicians/American College of Physicians
MRI Brain	70551, 70552, 70553	ACR; American Academy of Neurology; American Academy of Otolaryngology;U.S. Headache Consortium
MRI Lumbar Spine	72148, 72149, 72158	American Academy of Neurology; American College of Physicians/American Pain Society; ACR; North American Spine Society
MRI Knee	73721, 73722, 73723	ACR
MRI Shoulder	73221, 73222, 73223	ACR
SPECT-MPI	78451, 78452+	ACR; American College of Radiology

SOURCES: HCPCS codes: The Lewin Group, 2011b. Relevant guidelines: The Lewin Group, 2011a.
NOTES: Combined CT of abdomen and pelvis is also included. Of note, most HCPCS codes are Level 1 that are developed and published by the American Medical Association's CPT.
* New codes in 2011. + New codes in 2010.

While the evaluation focused on seven conveners, the composition of some of the conveners changed during the course of the evaluation. Two practices within a convener merged, one merging during the baseline period and the other after the baseline period. One convener remained a participant throughout MID, but beginning in December 2012 underwent a phased implementation of a new electronic health record system—a system that is incompatible with its DSS system for MID. Thus, the volume of DSS orders at that convener slowed over time as more providers ceased using DSSs.

1.3. General Approach and Requirements for the Demonstration

CMS established requirements for a DSS, and conveners implemented their own systems. As noted in Section 1.2, to maintain the confidentiality of focus group participants, we refer to conveners as Convener A, Convener B, etc. These names have been randomly assigned.

For the demonstration, all conveners, including clinicians or their designated staff, were to order advanced imaging associated with the MID demonstration using ROE and the DSS. The demonstration was implemented in two phases. During the first phase, sites that had been using paper-based orders changed their ordering to a ROE system on October 1, 2011, or soon after—or, if they were already using ROE, they updated the ROE system to be consistent with MID specifications. During the second phase—beginning April 1, 2012, or soon after—conveners

began implementing the DSS protocol so that ordering clinicians or practice staff could receive feedback about the appropriateness of their orders following order entry through the ROE system.

1.3.A. Decision Support Systems

Some DSSs are integrated with electronic medical record systems, while others function as Web-based or stand-alone software applications. The DSSs differ in their interfaces and functions, but all meet basic requirements stipulated by CMS. As described in the MID Convener Manual developed by CMS's MID implementation contractor, the Lewin Group, the following features characterize MID requirements:

- DSSs must be programmed with published clinical guidelines developed by approved medical specialty societies for MID advanced imaging procedures (see Table 1.2).
- Each imaging order should be associated with an appropriateness rating of "appropriate," "equivocal," "inappropriate," or "not covered by guideline."
- The appropriateness ratings should be based only on the approved clinical guidelines.
- Ordering clinicians should be shown the appropriateness rating and provided access to the source of the guideline concurrent with ordering the image when an order is "inappropriate" or "equivocal."
- Ordering clinicians should be shown more appropriate alternative tests or additional tests, if applicable.
- Ordering clinicians should have the prerogative to order any imaging procedure desired, regardless of its appropriateness rating, and the imaging procedures will be covered using existing Medicare coverage criteria.
- Prior to finalizing the order, clinicians must attest to placing the final order as originally specified, even if the order did not receive the highest appropriateness rating.

CMS and the implementation contractor accessed DSSs to ensure that they met the stipulations listed above and in the MID Convener Manual. Prior to implementation, conveners applied their DSSs to 30 to 50 test case scenarios developed by CMS and Lewin in consultation with medical societies. The test cases were used to assess the consistency of the DSS output with expected appropriateness ratings. However, Lewin considered this testing "limited" in its ability to ensure that DSSs are processing guidelines consistently because they cannot realistically test all the clinical scenarios that could be input into a DSS.[4]

The DSSs were implemented by the conveners with general guidance from Lewin. Conveners developed operational protocols for DSS users, and the protocols and plans were approved by CMS. Lewin provided technical assistance to the conveners, including four site visits per convener, a MID website, training webinars, and phone calls with conveners, both collectively and one on one. CMS provided payment to participants to support the MID conditional on participants meeting data reporting requirements. Payment for MID participation

[4] Discussed in phone meeting between CMS, Lewin, and RAND, January 31, 2012.

was not tied to a clinician's ordering volume, though clinicians were reimbursed for clinical care delivered as they would be if they were not participating in the MID. The payment amount was negotiated as part of the terms and conditions agreement signed at the initiation of the demonstration. The payments were given from CMS to the conveners, who were expected to distribute them to participating practices. To qualify for pay-for-reporting payments, practices had to meet a threshold for "completeness of reporting," calculated as the percentage of DSS records among billed MID advanced imaging procedures. In addition, participants received all other usual Medicare payments for services provided.

Conveners were responsible for providing additional aggregate feedback reports to their affiliated participating practices and physicians following guidance from Lewin. They had some flexibility in how they chose to present the data content, but were to provide these additional feedback reports on a quarterly basis.

Conveners varied in terms of how much information and training they provided to clinicians and staff on MID DSSs. They also varied with respect to the approach they used to structure the ROE and associated decision support. Conveners A, C, F, and G structured ROE by allowing ordering clinicians to enter patient clinical characteristics (e.g., age, gender, underlying disease processes, prior treatments) and the reason for the order (e.g., presence or absence of fever, pain, trauma). Each round of data entry had the potential to be followed by a series of queries to the ordering clinician or staff requesting additional information about the reason for the order. Conveners B, D, and E structured their ROE so that ordering clinicians would identify clinical patterns that mapped to variants (reasons for advanced imaging orders according to the guidelines) using drop-down menus available through the DSS interface.

Across conveners and even within conveners, there was substantial variation in the way clinicians and their practice staff members ordered advanced imaging. They varied with respect to whether clinically or administratively trained staff supplemented clinicians in placing advanced image orders—and if so, how the staff were involved. Some conveners developed a program whereby radiology order entry involved the identification of the guideline that explained the reason for the order. Other conveners developed a program whereby radiology order entry prompted those ordering advanced images to enter data that would allow an algorithm to assign the characteristics of the patient and their clinical conditions to a reason for order. Some conveners allowed clinicians to use DSSs and cancel or change orders using one integrated set of computer clicks. Within other convener practices, the DSS was not compatible with the electronic health record, which meant that each order cancellation or change was associated with an entirely new process for identifying the patient, ordering the image, and specifying the reasons for the order.

Each of the seven conveners recruited practices that covered a diverse set of clinical specialties (e.g., generalists, subspecialists). Conveners also differed in the extent to which individual specialties were represented in their participating clinics. This diversity in specialties of clinicians participating in the demonstration led to variations in the types of imaging

procedures ordered across conveners. Each convener implemented MID in a manner that reflected its own workflow. Conveners varied with respect to whether a DSS was applied to all patients or only to Medicare beneficiaries, and also varied with respect to urbanicity, size (number of patients, providers, and practices), prior exposure to radiology order entry, decision support, and use of electronic health records.

1.4. Approach to the Congressional Statute Questions

This last component of Section I introduces the remaining five sections.

Convener-Level Results Associated with Advanced Image Ordering

Section II, comprising Chapters 2, 3, and 4, presents convener-level results associated with advanced image ordering with DSSs for practices associated with the MID efforts. It addresses three statute questions:

- What were the rates of appropriate, uncertain, and inappropriate advanced imaging orders over the course of the demonstration? (Chapter 2)
- Were any patterns or trends evident in the appropriateness or inappropriateness of advanced imaging procedure orders? (Chapter 3)
- Is there a relationship between the appropriateness of advanced imaging procedure orders and imaging results? (Chapter 4)

Advanced Imaging Utilization Before and After MID

Section III, comprising Chapter 5, introduces analyses of advanced imaging utilization before and after MID was introduced in practices associated with MID compared with comparable control practices. It addresses the statute question:

- Were any national or regional patterns or trends evident in utilization of advanced imaging procedures?

Physician and Patient Experience with Appropriateness Criteria for Advanced Imaging

Section IV, comprising Chapters 6 and 7, addresses physician and patient experiences with exposure to advanced imaging appropriateness criteria. Statute questions include:

- How satisfied were physicians in the demonstration with being exposed to advanced imaging appropriateness criteria? (Chapter 6)
- How satisfied were Medicare patients in the demonstration with receiving an advanced imaging procedure after a physician was exposed to appropriateness criteria? (Chapter 7)

Statute Questions to Inform Recommendations for Future Decision Support

Section V, comprising Chapters 8 through 13, addresses six statute questions that can inform recommendations about future decision support.

- Was the system for determining appropriateness in the demonstration acceptable for identifying appropriate versus inappropriate advanced imaging orders? (Chapter 8)
- Would exposing physicians to advanced imaging appropriateness criteria at the time of order affect the volume of utilization? (Chapter 9)
- Is it advisable to expand the use of appropriateness criteria for ordering advanced imaging to a broader population of Medicare beneficiaries? (Chapter 10)
- If expanded to a broader population of Medicare beneficiaries, should physicians who demonstrate that their ordering patterns are consistently appropriate be exempt from requirements to consult appropriateness criteria? (Chapter 11)
- To what extent is live feedback on the appropriateness of advanced imaging orders from a decision support system better or worse than feedback reports to individual physicians or physician practices? (Chapter 12)
- In what ways can physicians be motivated—including financial incentives—to comply with ordering advanced imaging appropriately according to appropriateness criteria? (Chapter 13)

Conclusion

Section VI concludes the analytic component of this report (Chapter 14).

It is followed by two technical appendixes pertinent to analyses of DSSs and claims (Appendix A) and to clinician focus groups and surveys (Appendix B).

Section II: Convener-Level Results Associated with Advanced Image Ordering with Decision Support Systems for Practices Associated with the Medicare Imaging Demonstration

This section addresses three statute questions:

- What were the rates of appropriate, uncertain, and inappropriate advanced imaging orders over the course of the demonstration? (Chapter 2)
- Were any patterns or trends evident in the appropriateness or inappropriateness of advanced imaging procedure orders? (Chapter 3)
- Is there a relationship between the appropriateness of advanced imaging procedure orders and imaging results? (Chapter 4)

This page is intentionally blank.

2. Analysis of DSS Data

This chapter addresses the research question:

- What were the rates of appropriate, uncertain, and inappropriate advanced imaging orders over the course of the demonstration?

2.1. Introduction to the DSS Analysis

We begin with a description of the sample of beneficiaries for whom advanced imaging orders were placed using DSSs during the two-year demonstration period, and the clinicians who participated in the demonstration. We summarize the diverse specialties of these clinicians, as well as the relative volume and type of imaging procedures they ordered through DSSs—drawing contrasts between conveners and specialties where appropriate. We then present the results of three sets of unadjusted analyses; all results derived from regression models are presented in Chapter 3. First, we examined the likelihood of an order receiving an appropriateness rating by a DSS, the extent to which this rate changed between the baseline and intervention periods, and the degree of heterogeneity in these rates by procedure, convener, and clinician specialty. Second, we analyzed clinicians' decisionmaking following the receipt of feedback from the DSS. These analyses included the frequency with which a DSS provided recommendations for alternative procedures among orders that were initially rated inappropriate or equivocal or that were unrated, and the degree to which clinicians changed or canceled orders in response to feedback from the DSS. Because a clinician's decisionmaking appears to be influenced by the presence of alternative recommendations, all of these analyses are conducted separately for orders in which the DSS recommends an alternative procedure and orders for which the DSS provides no alternative. Finally, we summarize changes in the appropriateness of orders for imaging procedures between the baseline and intervention periods. We examine variation across procedures, conveners, and clinician specialties; and we assess the extent to which changes in appropriateness vary by a clinician's level of exposure to DSSs.

2.2. DSS Beneficiary and Clinician Samples

Orders for advanced imaging procedures were placed on behalf of nearly 80,000 unique Medicare beneficiaries who sought care from participating clinicians over the two-year period. In Table 2.1 we summarize the only demographic and utilization data that were collected through the DSS or that we could derive. Approximately 28 percent of beneficiaries were under the age of 65 at the start of the demonstration, while nearly 6 percent were over the age of 85. Just over 45 percent of beneficiaries were women. A single order for advanced imaging was placed on

behalf of two-thirds of beneficiaries by demonstration clinicians, however, clinicians ordered six or more procedures for nearly 4 percent of beneficiaries. Of note, while these results reflect *orders* for imaging procedures, they do not imply that all procedures were ultimately *rendered.* The disposition of all orders was not systematically tracked in a DSS so the volume of imaging procedures that were actually performed remains unknown. Because the goal of the DSS analyses was to capture changes in ordering behavior before and after implementation of appropriateness feedback, we included all orders for advanced imaging procedures in our analyses whether the procedures were rendered or not.

Table 2.1. Characteristics of Beneficiaries Included in the DSS Analysis

Characteristic	N (%)
Age (years)	
18–64	22,133 (27.8)
65–74	32,547 (40.9)
75–84	20,286 (25.5)
85+	4,705 (5.9)
Gender	
Female	43,179 (54.2)
Male	36,492 (45.8)
Number of DSS orders placed during demonstration	
1	52,407 (65.8)
2	15,204 (19.1)
3–5	9,109 (11.4)
6–9	2,167 (2.7)
10+	784 (1.0)
Total number of beneficiaries	**79,671 (100)**

A total of 5,128 clinicians placed orders for advanced imaging procedures using DSSs during the 24-month demonstration. Conveners differed widely in their approach for recruiting individual practices, with some conveners inviting mostly large, academic practices to participate in the demonstration, while others partnered with multiple community practices that tended to be smaller. As expected, the total number of clinicians who ordered through DSSs varied across conveners (Table 2.2). The largest convener (C) comprised 1,119 clinicians (22 percent of the total sample), while the smallest convener (E) had 376 participating clinicians (7 percent).

To enable stratified analyses by specialty, we assigned a single specialty to each ordering clinician using data from the National Plan and Provider Enumeration System (NPPES). For the small number of clinicians who could not be found in NPPES, we used specialty codes from claims.[5] We then used each clinician's assigned provider taxonomy code to create seven mutually exclusive categories: physician generalist, physician medical specialist, physician

[5] See Appendix A for more details on our methodology for assigning specialties to clinicians.

surgical specialist, nonphysician generalist, nonphysician specialist (including both medical and surgical specialists), and other.

Overall, the sample of clinicians was reasonably balanced across physician specialties. Approximately 29 percent of clinicians were generalists, while nearly one-third were medical subspecialists and just over 15 percent were surgical specialists. Nonphysicians, which included nurse practitioners and physicians assistants, comprised 17 percent of the sample—the vast majority of whom had training in a primary care specialty.

While each of the seven conveners, individually, sought to recruit a diverse set of clinical specialties, we found several notable differences in the distribution of specialties across conveners. For example, Convener E disproportionately recruited generalist physicians (61 percent of all clinicians recruited by Convener E) while clinicians affiliated with Conveners F and G were much more likely than the other five to have subspecialty training (59 percent and 63 percent of all recruited clinicians, respectively). Convener E recruited a disproportionately large share of cardiologists (12.5 percent of all clinicians affiliated with Convener E) but recruited no oncologists. All other conveners, with the exception of Convener A, recruited a sizable number of oncologists. Orthopedic surgeons represented a larger proportion of ordering clinicians at Convener A (5.5 percent) compared with other conveners, while only a single orthopedic surgeon affiliated with Convener F placed any orders through a DSS during the study period. The low participation of orthopedic surgeons was unexpected, since four of the 12 MID imaging procedures (CT lumbar spine, MRI lumbar spine, MRI shoulder, and MRI knee) are commonly ordered by orthopedic surgeons. Convener C differed notably from other conveners by having a much larger percentage of nonphysicians who placed orders through a DSS (just over 26 percent of all ordering clinicians). Convener C also had a substantially higher percentage of clinicians whose specialty we categorized as neither physicians nor nonphysicians (17 percent). Nearly 87 percent of these "other" clinicians were classified in NPPES as trainees. Convener D differed from all other conveners by having a very small percentage of orders placed by nonphysicians (less than 2 percent).

Overall, the rich diversity in specialties of clinicians participating in the demonstration allowed us to examine differences in the impact of DSSs between physicians and nonphysicians and also across different specialties. However, differences in specialties across conveners imply that individual conveners may be ordering different imaging procedures at very different rates for patients with very different clinical conditions. This level of heterogeneity, along with the known differences in design of each convener's DSS, suggests that an overall analysis might mask heterogeneity in the impact of the intervention. Thus, all analyses presented in this chapter are stratified by convener and specialty to account for these differences.

Table 2.2. Number of Participating Clinicians, by Convener and Specialty

Specialty	Convener							
	A	B	C	D	E	F	G	Overall
Physicians								
Generalists								
Internal Medicine	77 (16.2)	95 (9.7)	115 (10.3)	174 (23.9)	104 (27.7)	75 (16.5)	89 (8.9)	729 (14.2)
Family Medicine	70 (14.7)	169 (17.3)	127 (11.3)	72 (9.9)	111 (29.5)	2 (0.4)	23 (2.3)	574 (11.2)
Geriatric Medicine	8 (1.7)	12 (1.2)	3 (0.3)	8 (1.1)	11 (2.9)	10 (2.2)	14 (1.4)	66 (1.3)
Other	7 (1.5)	21 (2.2)	22 (2.0)	23 (3.2)	3 (0.8)	12 (2.6)	15 (1.5)	103 (2.0)
Total	162 (34)	297 (30.4)	267 (23.9)	277 (38)	229 (60.9)	99 (21.8)	141 (14.1)	1,472 (28.7)
Medical Specialists								
Cardiology	48 (10.1)	40 (4.1)	34 (3.0)	48 (6.6)	47 (12.5)	35 (7.7)	79 (7.9)	331 (6.5)
Oncology	5 (1.1)	64 (6.6)	30 (2.7)	36 (4.9)	0 (0.0)	34 (7.5)	87 (8.7)	256 (5.0)
Gastroenterology	4 (0.8)	26 (2.7)	20 (1.8)	18 (2.5)	15 (4.0)	24 (5.3)	40 (4.0)	147 (2.9)
Pulmonology	21 (4.4)	24 (2.5)	9 (0.8)	21 (2.9)	3 (0.8)	11 (2.4)	35 (3.5)	124 (2.4)
Neurology	17 (3.6)	32 (3.3)	26 (2.3)	51 (7.0)	5 (1.3)	31 (6.8)	60 (6.0)	222 (4.3)
Other	40 (8.4)	94 (9.6)	95 (8.5)	114 (15.6)	25 (6.6)	57 (12.5)	162 (16.2)	587 (11.4)
Total	135 (28.4)	280 (28.7)	214 (19.1)	288 (39.5)	95 (25.3)	192 (42.2)	463 (46.4)	1667 (32.5)
Surgical Specialists								
Orthopedic Surgery	26 (5.5)	40 (4.1)	25 (2.2)	20 (2.7)	7 (1.9)	1 (0.2)	27 (2.7)	146 (2.8)
Urology	13 (2.7)	15 (1.5)	14 (1.3)	13 (1.8)	1 (0.3)	11 (2.4)	17 (1.7)	84 (1.6)
Otolaryngology	4 (0.8)	15 (1.5)	19 (1.7)	12 (1.6)	0 (0.0)	9 (2.0)	23 (2.3)	82 (1.6)
Thoracic Surgery	0 (0.0)	5 (0.5)	10 (0.9)	5 (0.7)	3 (0.8)	1 (0.2)	10 (1.0)	34 (0.7)
Other	29 (6.1)	82 (8.4)	88 (7.9)	83 (11.4)	6 (1.6)	54 (11.9)	89 (8.9)	431 (8.4)
Total	72 (15.1)	157 (16.1)	156 (13.9)	133 (18.2)	17 (4.5)	76 (16.7)	166 (16.6)	777 (15.2)
Nonphysicians								
Generalists	66 (13.9)	147 (15.1)	249 (22.3)	6 (0.8)	16 (4.3)	41 (9.0)	139 (13.9)	664 (12.9)
Specialists	17 (3.6)	66 (6.8)	45 (4.0)	4 (0.5)	9 (2.4)	12 (2.6)	70 (7.0)	223 (4.3)
Other	24 (5.0)	29 (3.0)	188 (16.8)	21 (2.9)	10 (2.7)	35 (7.7)	18 (1.8)	325 (6.3)
Total	**476**	**976**	**1,119**	**729**	**376**	**455**	**997**	**5,128**

NOTE: Cell values represent counts of unique clinicians; percentages are in parentheses. All clinicians who placed at least one order through a DSS during the 24-month demonstration period were considered to be participating in the demonstration.

2.3. DSS Image Order Sample

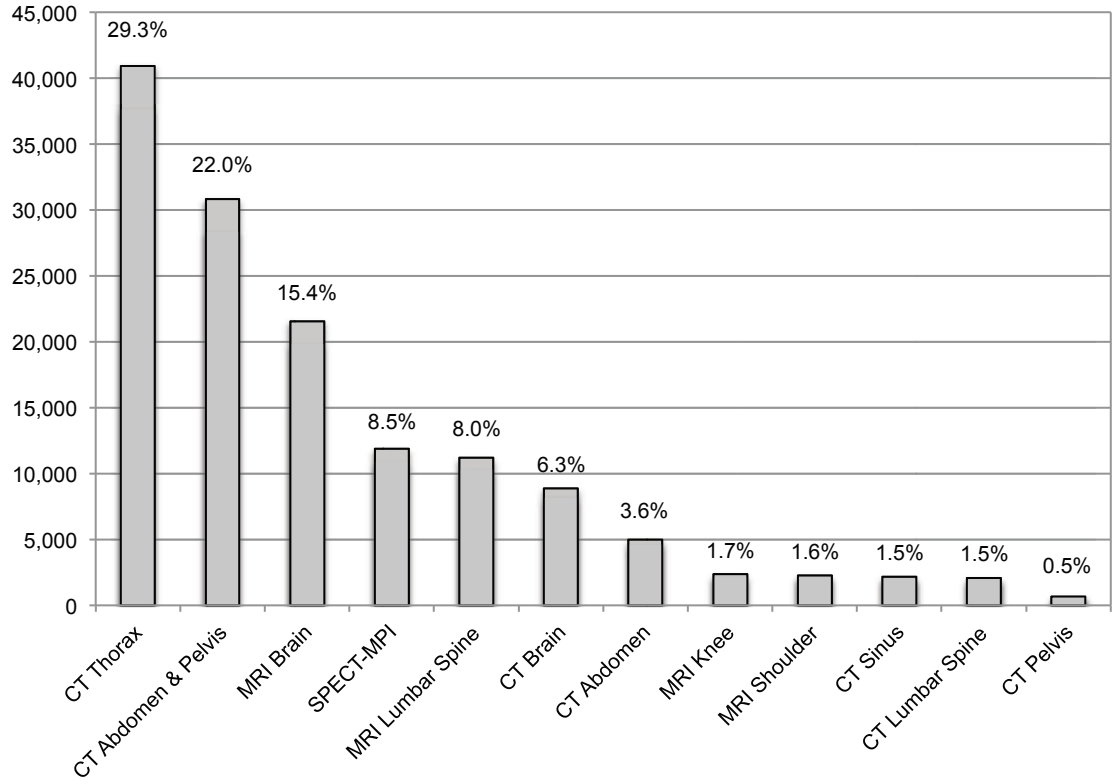
Participating clinicians placed a total of 139,757 orders using DSSs over the 24-month demonstration. Figure 2.1 displays the distribution of "initial" orders.[6] Initial orders refer to the order each clinician enters into a DSS before being presented with feedback on the appropriateness of the order, which occurs during the intervention period only. Thus, these

[6] Combination orders for CT abdomen and CT pelvis were tracked as a unique type of order within all conveners' DSSs because the combination procedure is ordered frequently and is appropriate in many cases. We treat the combined procedure as a single procedure in all analyses.

procedures do not represent the "final" order if the clinician changes or cancels the order in response to receiving appropriateness feedback.

Figure 2.1 displays the frequency of initial orders by procedure. Just over 29 percent of all orders were for CT thorax, followed by the combination procedure CT abdomen/CT pelvis (22 percent of all orders). The volume of orders for CT sinus, MRI shoulder, and MRI knee appear quite low, given the number of clinicians participating in MID who commonly order these procedures. For example, 82 otolaryngologists participated in MID and these clinicians ordered 2,158 CT sinus procedures over 24 months, which is a rate of approximately 1.1 orders per clinician per month. Similarly, the 146 orthopedic surgeons who placed at least one order in a DSS had an average ordering rate of 0.6 MRI shoulder procedures per clinician per month and 0.7 MRI knee procedures per clinician per month. These rates are much lower than expected—particularly for clinicians affiliated with large academic practices.

Figure 2.1. Volume of Initial Orders, by Procedure



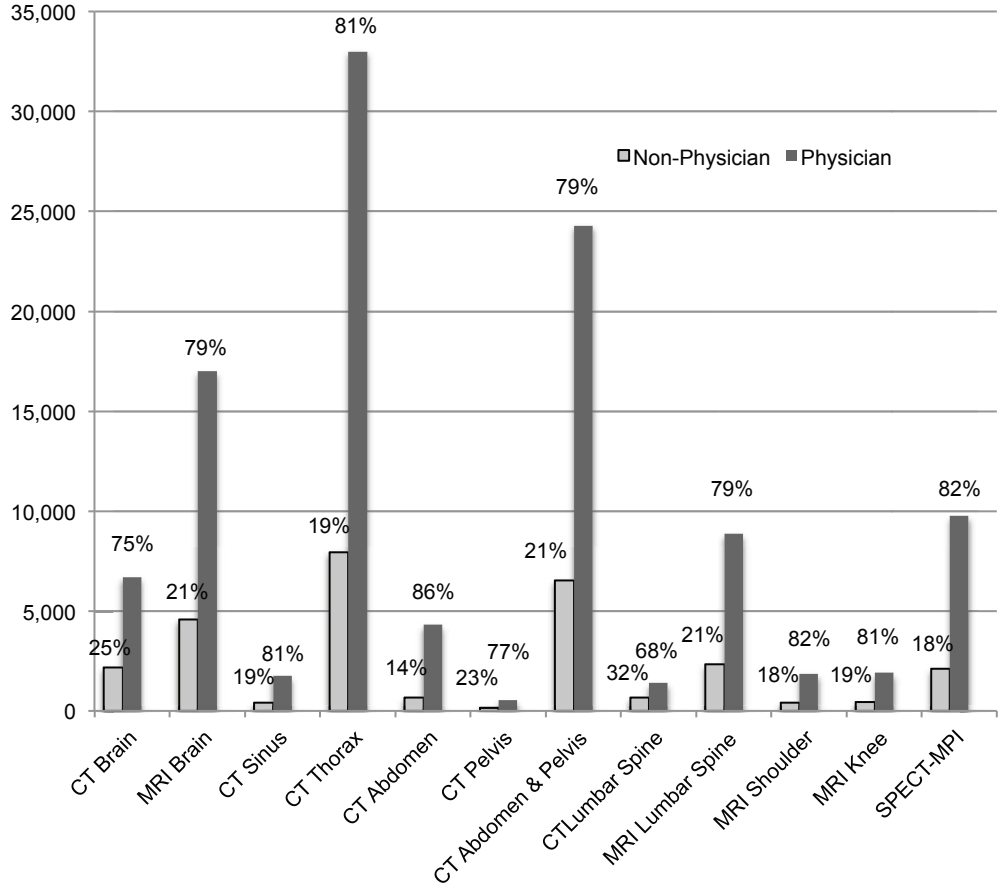
NOTE: The total sample size is 139,757 "initial" orders. "Initial" refers to a clinician's initial selection of an imaging procedure prior to receiving appropriateness feedback (available during the 18-month intervention period only).

We then examined the extent to which orders for each procedure were placed by physicians relative to nonphysicians. Each category of clinicians may be influenced by DSSs in different ways. For example, *physicians* might be more likely to ignore DSS feedback if they have significant concerns about the quality of the guidelines upon which the appropriateness ratings

are based. By contrast, *nonphysicians* might be more receptive to DSS feedback if they have less prior knowledge of imaging guidelines or if they are less likely to question their validity. Both types of clinicians might be more or less likely to change orders in response to DSS feedback depending on the use of "protocoling" of advanced image orders in their site.[7] We looked for evidence to support each of these hypotheses.

In general, imaging procedures are far more likely to be ordered through DSSs by physicians—ranging from 75 percent to 86 percent of orders across procedures. Compared to other procedures, orders for CT lumbar spine are slightly more likely to be ordered by nonphysicians.

Figure 2.2. Volume of Initial Orders, by Procedure and Type of Clinician

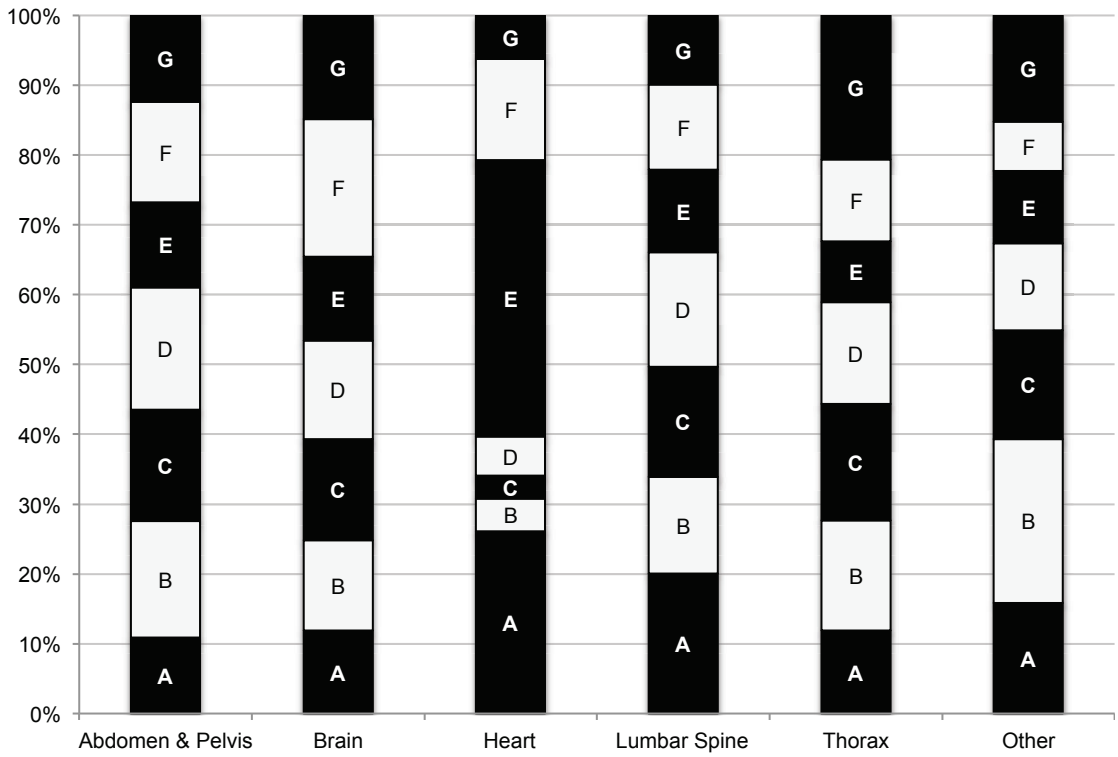


[7] "Protocoling" refers to the systematic review of all radiology orders by radiologists. Some conveners have a policy of "protocoling" all advanced imaging orders.

Variability in the distribution of clinician specialties across conveners (as illustrated in Table 2.2) contributed to substantial differences in the type of procedures ordered by clinicians affiliated with each convener (Figure 2.3). For example, Conveners A and E (where cardiologists are most highly represented) dominate the sample of orders for SPECT-MPI—an imaging procedure for the heart. In fact, conveners with the highest and lowest rate of SPECT-MPI ordering differ by a factor of more than 12. Ordering rates for CT thorax also differ notably across conveners, from a low of 16 percent of all imaging procedures for Convener E to a high of 39 percent for Convener G.

Another source of variability that exists between conveners is the rate at which MRI or CT imaging modalities are ordered for the same body system, which is not displayed in Figure 2.3. For example, MRI and CT procedures of the brain are ordered at roughly equal rates for Conveners A and E, in a 2:1 ratio for Conveners B, C, D, and F, and approximately a 5:1 ratio for Convener G. Similarly, MRI procedures of the lumbar spine, when compared to CT, range from a 3:1 ratio to a 10:1 ratio across conveners compared with CT procedures. Whether these differences reflect variability in practice patterns or differences in clinicians' access to different types of imaging equipment is something that we explored through focus groups.

Figure 2.3. Distribution of Initial Orders for MID Imaging Procedures, by Body System and Convener



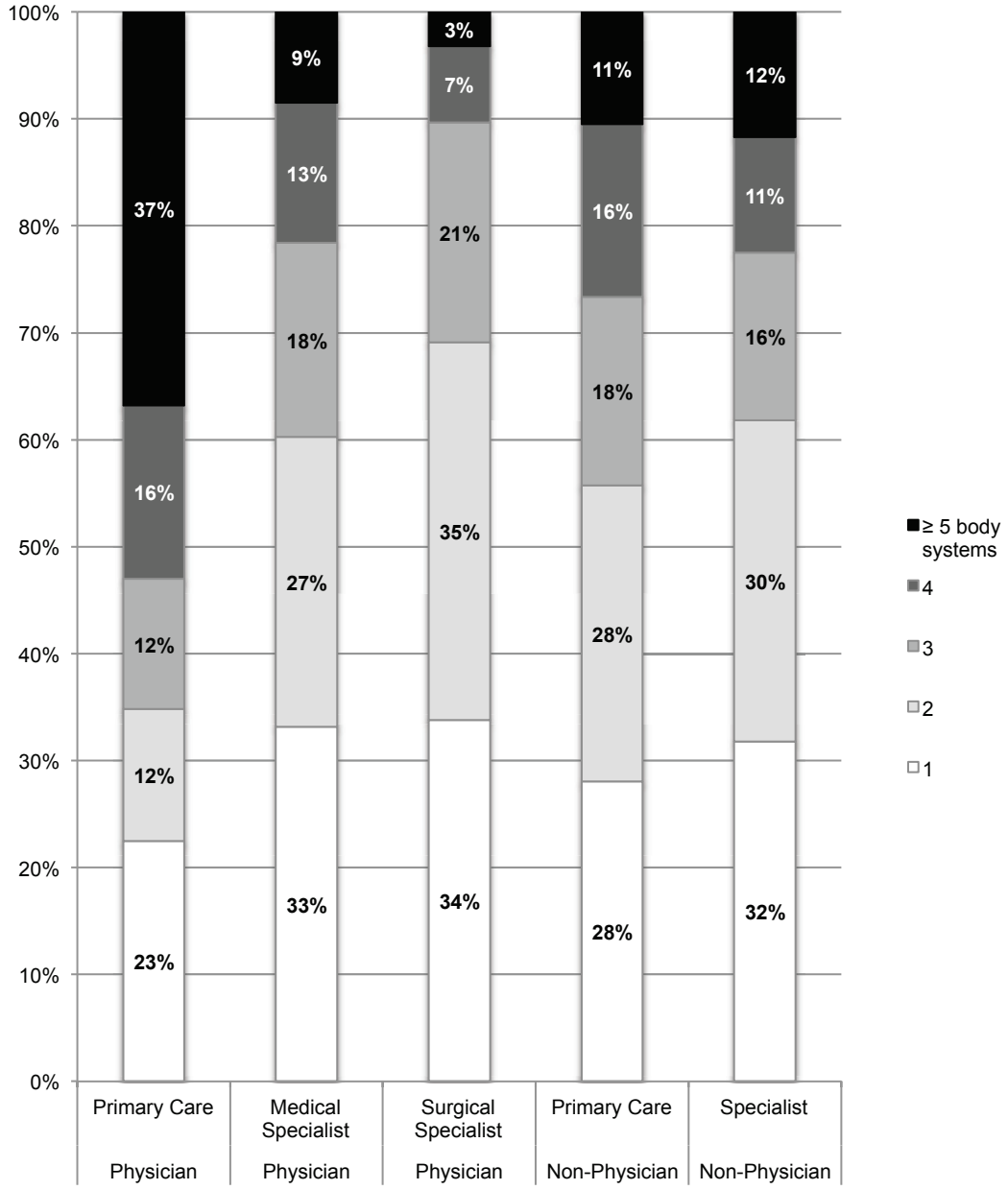
NOTE: The figure displays the percentage of orders placed by each convener for each of six categories of body systems. Brain and lumbar spine includes both CT and MRI procedures. Heart includes SPECT-MPI only. "Other" includes orders for CT knee, CT shoulder, and CT sinus. Ordering rates for CT knee, shoulder, and sinus differed very little across conveners. "Initial" orders are shown. Clinicians might have modified their initial order in response to DSS feedback. Letters refer to Conveners A–G.

The potential for DSSs to affect ordering patterns will depend highly on the frequency with which individual clinicians order advanced imaging procedures. We hypothesized that ordering volume and thus "exposure" to the DSS intervention will be related to a clinician's specialty. In particular, we hypothesized that specialists would be more likely to order a large volume of procedures through DSSs for a relatively narrow set of procedures, while generalists would have lower imaging volumes but would order a broader set of procedures.

Our analyses indicate that generalist physicians were far more likely to order images for a larger number of body systems compared to subspecialists (Figure 2.4). Nearly 37 percent of primary care physicians ordered images on five or more body systems, compared to only 9 percent and 3 percent of medical and surgical subspecialists, respectively. The breadth of orders placed by nonphysicians was comparable to that of physician subspecialists.

Differences in ordering patterns such as these may influence the effectiveness of a DSS in a number of ways. It may be "easier" for a DSS to influence the ordering patterns of specialty clinicians because feedback may be given repeatedly for the same procedures ordered for a relatively narrow set of clinical indications. It might also be the case that generalists might rely more heavily on DSSs given the wide body of literature on the appropriateness of imaging for five or more body systems. Analyses presented in this chapter are stratified by specialty to account for these systematic differences in ordering patterns across specialties.

Figure 2.4 Number of Body Systems Imaged by Clinician Specialty



NOTE: The figure displays the distribution of the count of body systems for which each clinician placed at least one order through a DSS during the demonstration.

Aside from heterogeneity in the *type* of imaging procedures ordered by each clinician, the total *volume* of orders placed by each clinician may also moderate the impact of DSSs. Approximately two-thirds of clinicians in our sample placed fewer than 20 orders (Table 2.3) Just over half of clinicians (52 percent) placed ten or fewer orders. Only 15 percent of clinicians ordered 50 or more imaging procedures through DSSs. To put these numbers into context, 50 orders represents approximately one order entered into a DSS biweekly over the course of the

two-year demonstration (for clinicians participating for the entire period). For no convener did more than 20 percent of clinicians order more than 50 advanced imaging procedures.

Clinicians' ordering volumes differed widely across specialties, as would be expected. Specialists had substantially higher ordering volumes than generalists. A total of 21.4 percent of participating medical specialists and 15.7 percent of surgical subspecialists ordered 50 or more imaging procedures over the course of the demonstration, compared to only 9.3 percent of generalist physicians. Physician subspecialists were far more likely to order more than 100 images (10.4 percent and 6.6 percent for medical and surgical specialists, compared to 1.4 percent for generalists). Nonphysicians (e.g., nurse practitioners, physician assistants) had ordering volumes that were slightly higher than those of generalist physicians.

Analyses presented later in this chapter use a clinician's ordering volume as a proxy for exposure to the DSS intervention. We then compare changes in appropriateness of ordering between the baseline and intervention periods across different levels of "exposure."

Table 2.3. Clinician-Level Ordering Volume, by Clinician Specialty

Clinician Specialty	<20 orders (%)	20-49 orders (%)	50-199 orders (%)	>=100 orders (%)
Total number of clinicians, n (%)	*3,427 (66.8)*	*932 (18.2)*	*460 (9)*	*309 (6)*
Physician Generalist				
Internal Medicine	61.6	26.2	10.3	1.9
Family Medicine	78.2	16.0	5.1	0.7
Geriatric Medicine	74.2	15.2	7.6	3.0
Other	88.3	3.9	6.8	1.0
Total	70.5	20.2	7.9	1.4
Physician: Medical Specialist				
Cardiology	54.7	24.2	16.3	4.8
Oncology	29.7	18.4	13.3	38.7
Gastroenterology	72.8	21.1	4.1	2.0
Pulmonology	37.1	25.0	21.0	16.9
Neurology	50.5	23.0	15.3	11.3
Other	80.9	12.6	4.9	1.5
Total	59.8	18.8	11.0	10.4
Physician: Surgical Specialist				
Orthopedic Surgery	71.2	21.2	5.5	2.1
Urology	36.9	26.2	19.0	17.9
Otolaryngology	48.8	30.5	15.9	4.9
Thoracic Surgery	55.9	14.7	8.8	20.6
Other	71.5	16.2	7.2	5.1
Total	64.6	19.7	9.1	6.6
Nonphysician				
Generalist	68.5	16.3	8.4	6.8
Specialist	71.3	15.2	9.4	4.0
Other	84.9	8.0	4.0	3.1

2.4. Unadjusted Results—Likelihood of Orders Receiving an Appropriateness Rating

CMS selected the 12 imaging procedures for the demonstration (described above) mainly because they are among the advanced imaging procedures that clinicians most commonly order for their Medicare patients. Analysis of the indications frequently associated with these procedures conducted prior to the demonstration by the implementation contractor, The Lewin Group, indicated that between 53 percent and 89 percent of orders for these procedures should be successfully matched to a guideline and therefore be assigned an appropriateness rating by conveners' DSSs during the demonstration. However, to encourage innovative models, CMS gave conveners the flexibility to determine how patient information entered in DSSs would be linked to a specific guideline. Two broad methods were used. Four conveners structured DSSs by allowing ordering clinicians to enter patient clinical characteristics (e.g., age, gender, underlying disease processes, prior treatments) and the reason for the order (e.g., presence or absence of fever, pain, trauma). Each round of data entry had the potential to be followed by a series of queries to the ordering clinician or staff requesting additional information about the reason for the order. Three conveners structured their DSSs so that ordering clinicians would identify clinical patterns that mapped to variants (reasons for advanced imaging orders according to the guidelines) using drop-down menus available through the DSS interface.

The implementation contractor tested the validity of each convener's mapping algorithm using a small number of test cases; however, the algorithm was considered the intellectual property of each convener and was not made available to RAND.

The actual percentage of orders receiving an appropriateness rating in both the baseline and intervention periods was much lower than anticipated (see Table 2.4). In the baseline period, during which clinicians received neither feedback on whether the order was associated with an available guideline nor the appropriateness of the order, we anticipated observing relatively low rates of orders with appropriateness ratings because clinicians were becoming familiar with the order entry process. We hypothesized that the percentage of orders that were linked to a guideline and rated would be higher during the intervention period after clinicians had mastered order entry.

Contrary to what we expected, the vast majority of orders across both baseline and intervention periods were not rated—nearly two-thirds of all orders. This pattern was consistent across procedures; DSSs were unable to assign appropriateness ratings to more than half of orders for nine of 12 MID procedures. For most procedures, the percentage of orders successfully mapped to a guideline was 20 to 40 points lower than expected. Moreover, rather than observing an increase in the percentage of rated orders between the baseline and intervention periods, we observed an overall decrease (2.8 percentage points) and decreases for all but two procedures (CT pelvis and CT abdomen), which increased by 3 and 6 percentage

points, respectively. The largest decrease in the percentage of rated orders was for orders for CT and MRI of the lumbar spine, which decreased by 8 and 15 percentage points, respectively.

Table 2.4. Change in Percentage of Orders Receiving Appropriateness Ratings Between Baseline and Demonstration Periods, by MID Imaging Procedure

		Baseline Period		Intervention Period		
Order	Anticipated Percentage of Rated Orders*	Number of Rated Orders	Percentage Rated	Number of Rated Orders	Percentage Rated	Change in Percentage Rated (Intervention-Baseline)
CT Brain	64	766	43.6	2,681	37.7	–5.9
MRI Brain	57	1,310	37.0	6,156	34.2	–2.8
CT Sinus	72	179	49.4	778	43.3	–6.1
CT Thorax	66	2,352	37.1	12,836	37.1	0.0
CT Abdomen	58	104	10.9	698	17.2	6.3
CT Pelvis	53	6	5.9	55	9.3	3.4
CT Abdomen and Pelvis	-	819	18.9	5,164	19.5	0.6
CT Lumbar Spine	83	160	43.5	609	35.8	–7.7
MRI Lumbar Spine	84	1,039	53.5	3,605	39.0	–14.5
MRI Shoulder	73	200	54.9	957	50.4	–4.5
MRI Knee	78	179	41.0	713	37.3	–3.7
SPECT-MPI	89	1,231	66.3	6,284	62.8	–3.5
Overall		8,345	37.3	40,536	34.5	–2.8

*SOURCE: Share of procedures covered by guidelines: The Lewin Group, 2010
NOTES: A “rated” order denotes an order for which the DSS was able to match the procedure and associated indications to a guideline selected by CMS for the demonstration.

To better understand these patterns, we examined the extent to which changes in the percentage of rated orders differed by convener and, within convener, by a clinician’s specialty. The magnitude of the differences we observed across conveners was substantial. Focusing first on the baseline period, the percentage of orders that were successfully rated by DSSs ranged from as little as 17 percent for Convener D to a high of 58 percent for Convener A (see Table 2.5). Changes in the level of rated orders followed different trajectories for different conveners. For three conveners (A, B, and D) the percentage of rated orders remained relatively unchanged between the two periods. For two conveners (F, and G), the percentage of orders that were assigned a rating dropped moderately (by 4 and 6 percentage points, respectively). In Convener C, however, the percentage of rated orders dropped by a staggering 24 points. Only a single convener (E) exhibited a substantial increase in the percentage of rated images over time (18 percentage points).

Table 2.5. Percentage of Orders Receiving Appropriateness Ratings, by Demonstration Period and Convener

Convener	Baseline Period		Intervention Period		Change in Percentage Rated (Intervention-Baseline)
	Number of Rated Orders	Percentage Rated	Number of Rated Orders	Percentage Rated	
A	1,521	57.9	5,814	59.7	1.8
B	624	20.5	3,058	20.1	-0.4
C	1,947	50.7	6,298	26.5	-24.2
D	914	16.6	2,847	18.6	2.0
E	957	52.3	5,789	69.8	17.5
F	919	45.2	5,448	41.0	-4.2
G	1,463	41.9	11,282	35.6	-6.3
Overall	**8,345**	**37.3**	**40,536**	**34.5**	**-2.8**

NOTES: A "rated" image denotes an order for which the DSS was able to match the procedure and associated indications to a guideline selected by CMS for the demonstration

Our analyses that stratified by both convener and specialty found somewhat mixed results. The conveners with the largest increases and decreases in the percentage of rated orders (Conveners E and C), had consistent patterns across all specialty categories (see Table 2.6). For five of the six remaining conveners, generalist physicians were less likely to receive ratings for their orders, and in the case of Convener G, substantially less likely. Among medical and surgical specialists the percentage of rated orders increased for some conveners and decreased for others, and even within conveners medical and surgical specialists often exhibited different patterns. While orders placed by nonphysicians were typically much less likely to be rated over time, nonphysicians ,make up a much smaller share of the sample of ordering clinicians compared with physicians.

Table 2.6. Change in Percentage of Orders Receiving Appropriateness Ratings Between Baseline and Demonstration Periods, by Clinician Specialty and Convener

Convener	Physicians			Non-physicians		Overall
	Generalist	Medical Specialist	Surgical Specialist	Generalist	Medical or Surgical Specialist	
A	-3.2	7.1	-2.8	-10.1	1.7	1.8
B	-6.5	-0.5	2.6	3.2	5.9	-0.4
C	-26.5	-23.3	-18.3	-23.9	-25.7	-24.2
D	0.2	2.5	5.6	-	-	2.0
E	29.9	7.0	-	-	-	17.5
F	-2.4	-4.3	-1.0	-10.9	-11.2	-4.2
G	-11.1	-6.3	-0.1	-9.3	-2.9	-6.3
Overall	**0.3**	**-1.2**	**1.9**	**-12.7**	**-5.4**	**-2.8**

NOTE: A minimum of 100 orders within specialty categories was required for this analysis.

The reasons for the high percentage of unrated orders is unclear. One possibility is that there were technical problems in the way in which each DSS mapped an order to a guideline. For example, clinicians from several conveners commonly ordered multiple procedures for the same patient at the same time. The most common procedure was a CT thorax combined with a CT abdomen/pelvis. The ability of each convener's DSS to accommodate combination orders is unclear; however, RAND is planning additional outreach to conveners to better understand whether combination orders might have contributed to the high percentage of unrated orders. As mentioned previously, RAND does not have access to each convener's mapping logic to independently test this theory. It may also be that the patients seen by clinicians participating in the demonstration simply do not have profiles that fit existing guidelines—for example, because the patient's condition has attributes that are not described in the guideline, or the guidelines do not account for a patient's other comorbidities.

The lack of an increase in rated orders over the course of the demonstration raises some concerns. The large reduction we observed for Convener C can be linked to an internal decision made by the convener lead to scale back the use of DSSs during the intervention period (See Figure 2.10, Panel C). Clinicians within all conveners may also have learned that the time required to engage with DSSs could be reduced through shortcuts, such as writing in indications or selecting nonspecific indications, both of which would prevent DSSs from matching an order to a guideline.[8] These observations are based on anecdotal reports of focus group participants; we have no empirical evidence to support this theory. The increase in unrated orders may have offset an increase in rated orders among clinicians who mastered order entry.

Finally, the persistence of high levels of unrated orders during the intervention period might be explained by the fact that certain procedures that were unrated in the baseline period were ordered repeatedly for the same patient during the intervention period. High levels of repeated orders are supported by Table 2.1, which shows that one-third of patients had multiple orders placed on their behalf. Two procedures that are among the least likely to be rated in the baseline period, CT thorax and CT abdomen/pelvis (which are ordered disproportionately by oncologists in our sample), are also the two procedures that are most likely to be ordered repeatedly over time to monitor disease progression. Thus, the high prevalence of unrated orders may be associated with a large volume of repeat orders for which the initial order was unrated.

2.5. Clinician Decisionmaking Following DSS Feedback

The premise for providing appropriateness feedback through DSSs is to guide clinicians toward more clinically appropriate procedures based on the scientific evidence as codified in

[8] Conveners' DSSs varied in the way in which indications for each order were entered. For some conveners, clinicians were able to select a single reason for the order, while other conveners used sophisticated algorithms based on decision trees to determine the reason for the order based on clinicians' responses to queries generated by the DSS. In the former case, clinicians were much more likely to avoid a rating by a DSS by selecting "other reason" for the order (i.e., the reason was not listed in the drop down menu available to the ordering clinician).

clinical guidelines. Accordingly, we hypothesized that the availability of an alternative procedure with a superior level of appropriateness would be a key factor influencing a clinician's decision to modify an order that initially received an inappropriate or equivocal rating.

Table 2.7 displays the percentage of orders for which DSSs provided at least one alternative order with an equal or higher appropriateness score. Among orders initially rated inappropriate, DSSs recommended alternative procedures for slightly less than 40 percent of orders, although there was substantial heterogeneity across procedures. For example, clinicians were prompted to switch to an alternative order for more than three-quarters of all inappropriate orders for CT abdomen, while less than 2 percent of inappropriate SPECT-MPI orders were associated with alternatives.

The high percentage of inappropriate orders for which no alternative is recommended might be explained by the fact that in no cases did a convener's DSS provide a recommendation to "perform no imaging." For example, 99 percent of inappropriate SPECT-MPI orders most likely did not trigger an alternative order because a heart function test was simply not indicated for the patient. In addition, the alternatives reported by the DSS were only alternative *imaging* procedures and did not include nonradiologic options; in many cases the appropriate clinical action may be to avoid imaging altogether. By contrast, DSSs provided recommendations for alternative orders for the majority of orders initially rated equivocal—nearly 80 percent of such orders.

Table 2.7 Percentage of Orders for Which DSSs Provided at Least One Alternative Order, by Procedure

	"Inappropriate" Initial Orders				"Equivocal" Initial Orders		
	Initial Orders	Orders with ≥1 Alternative			Initial Orders	Orders with ≥1 Alternative	
Order	N	N	%	Order	N	N	%
CT Abdomen	50	39	78.0	CT Lumbar Spine	284	279	98.2
CT Lumbar Spine	53	36	67.9	MRI Lumbar Spine	436	428	98.2
MRI Shoulder	163	100	61.3	CT Abdomen	98	95	96.9
CT Brain	100	60	60.0	CT Brain	947	896	94.6
CT Sinus	13	7	53.8	MRI Brain	280	248	88.6
MRI Lumbar Spine	281	148	52.7	CT Abdomen and Pelvis	1,172	1,023	87.3
CT Pelvis	18	9	50.0	MRI Knee	29	22	75.9
MRI Knee	185	84	45.4	CT Sinus	36	27	75.0
CT Thorax	872	381	43.7	CT Thorax	1,573	1,174	74.6
CT Abdomen and Pelvis	446	170	38.1	CT Pelvis	21	13	61.9
MRI Brain	324	80	24.7	MRI Shoulder	13	7	53.8
SPECT-MPI	403	6	1.5	SPECT-MPI	477	14	2.9
Total	**2,908**	**1,120**	**38.5**	**Total**	**5,366**	**4,226**	**78.8**

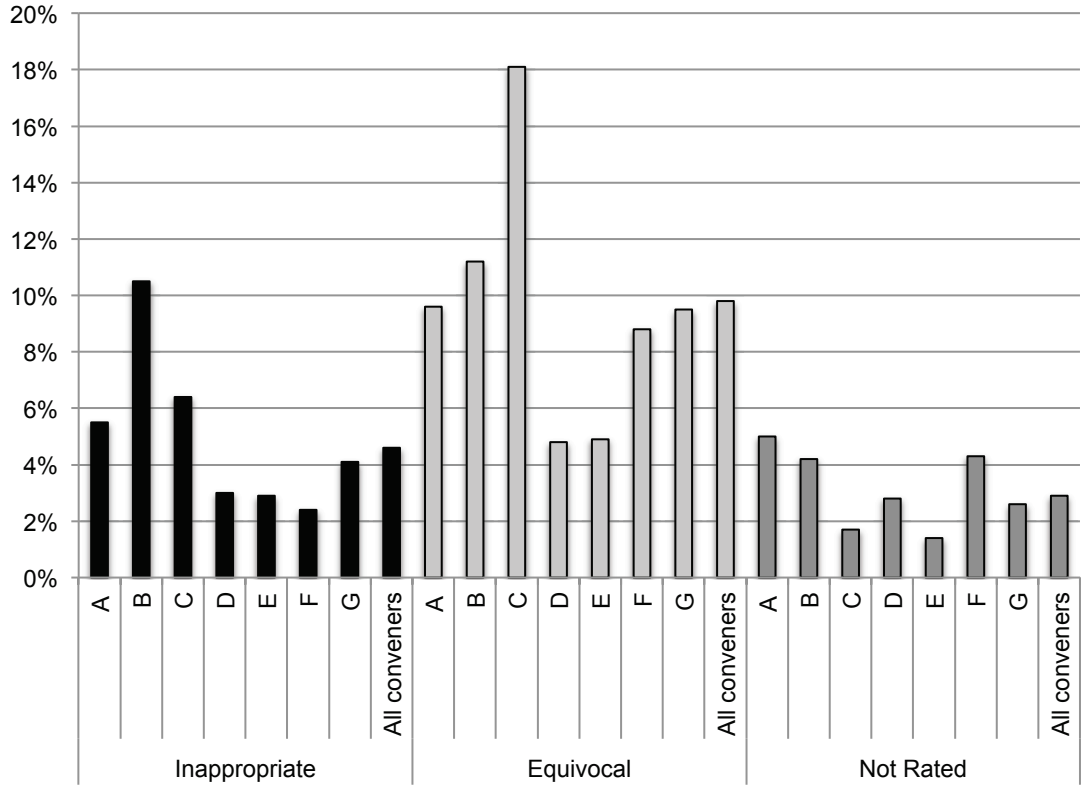
NOTE: Data from the intervention period only are displayed. (Recommendations for alternative procedures are not provided by DSSs during the baseline period). Percentages are sorted in descending order.

Each convener's DSS captured changes to each order during the intervention period (when clinicians received feedback on the appropriateness of their order and, if available, recommendations for alternative orders). If the DSSs were working as intended, we would expect a large proportion of orders initially rated inappropriate to be changed to a more appropriate order or canceled during the intervention period when clinicians receive feedback.

However, Figure 2.5 illustrates the extremely low rates of orders that are changed by clinicians during the intervention period—even when the initial order triggered an inappropriate rating. No convener changed modalities (CT to MRI, MRI to CT, or advanced imaging to a non-MID procedure) or changed contrast status more than 11 percent of the time when the initial order was rated inappropriate. Conveners were more likely to change their initial order when receiving an equivocal rating, but only marginally so. The reluctance of clinicians to change their initial order in these scenarios may reflect clinicians' lack of certainty of the validity of the linkage between the DSS and the reason for order as intended by the clinician or the validity of the guideline upon which the inappropriate rating was based. Alternatively, the use of protocoling within many of the conveners' delivery systems may have reduced the incentive for clinicians to change their order—particularly if they viewed their local radiologist's input as more trustworthy than a national guideline.

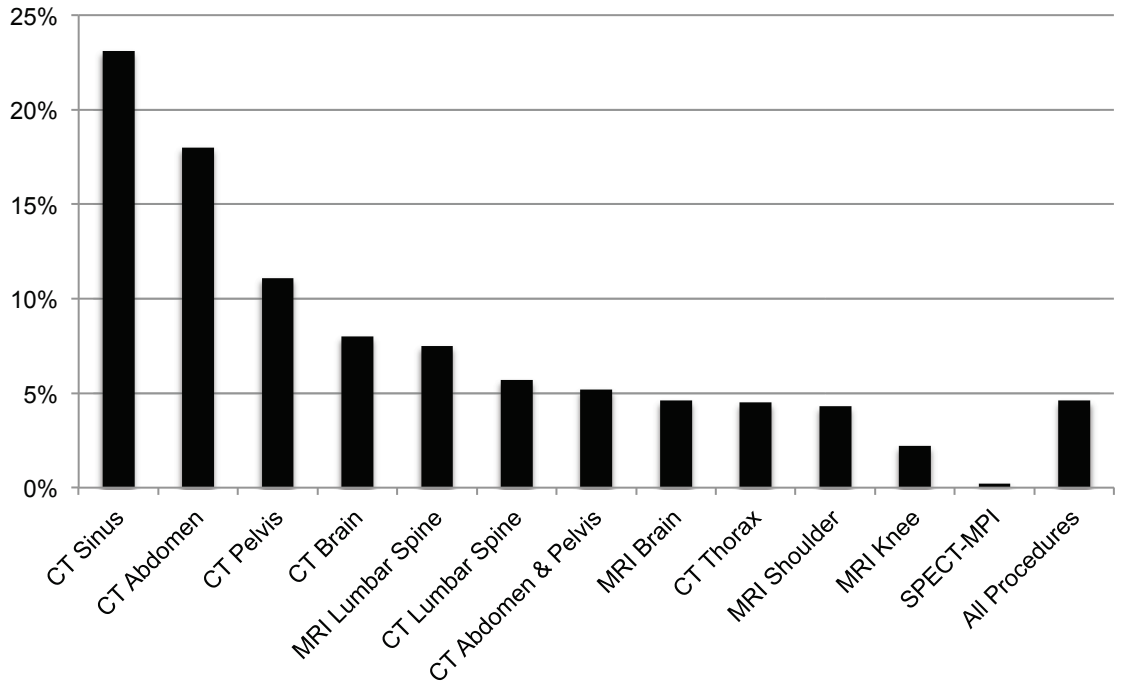
Focusing on only the inappropriate orders, we examined whether or not the rate of changes to initial orders varied across procedures (Figure 2.6). The frequency with which clinicians changed orders that were initially rated inappropriate was remarkably consistent across procedures. With few exceptions, clinicians changed their orders approximately 4 to 8 percent of the time for most procedures. The rates reported for CT sinus, CT abdomen, and CT pelvis were based on very small sample sizes, so those rates should be interpreted with caution. Future analyses planned by RAND will explore the clinical conditions associated with the highest rates of inappropriate orders (using data we have recently acquired). We will then better understand which orders are associated with alternatives, and among those that are, the degree to which the presence of an alternative induces a clinician to change his or her order.

Figure 2.5. Percentage of Orders Changed After the Initial Order, by Appropriateness Rating, Convener, and Intervention Period



NOTE: Letters denote Conveners A–G. The results are stratified by the appropriateness rating assigned to the initial order

Figure 2.6. Percentage of Inappropriate Orders Changed After the Initial Order, by Procedure, Intervention Period



NOTE: Only initial orders rated inappropriate are displayed. All conveners are included.

Rather than changing their initial order, clinicians might decide to cancel their order after receiving an inappropriate or equivocal rating. We expect the highest rates of cancellations when orders are rated either inappropriate or equivocal or when no alternatives are available from the DSS. Clinicians might also decide to cancel their orders if the DSS recommends alternative procedures but the clinician does not agree with the alternatives. We examined the rates at which clinicians canceled orders during the intervention period and found that rates of cancellation were much lower than rates of changes with few exceptions (see Figure 2.7). Convener D was associated with a particularly high rate of cancellation—particularly for orders initially rated inappropriate. Clinicians associated with all other conveners rarely canceled orders—even when the initial order received an inappropriate rating—and Conveners C and G never canceled orders. CT lumbar spine and MRI knee were the two procedures most commonly canceled by clinicians after receiving inappropriate ratings (see Figure 2.8).

Figure 2.7. Percentage of Orders Canceled After the Initial Order, Intervention Period

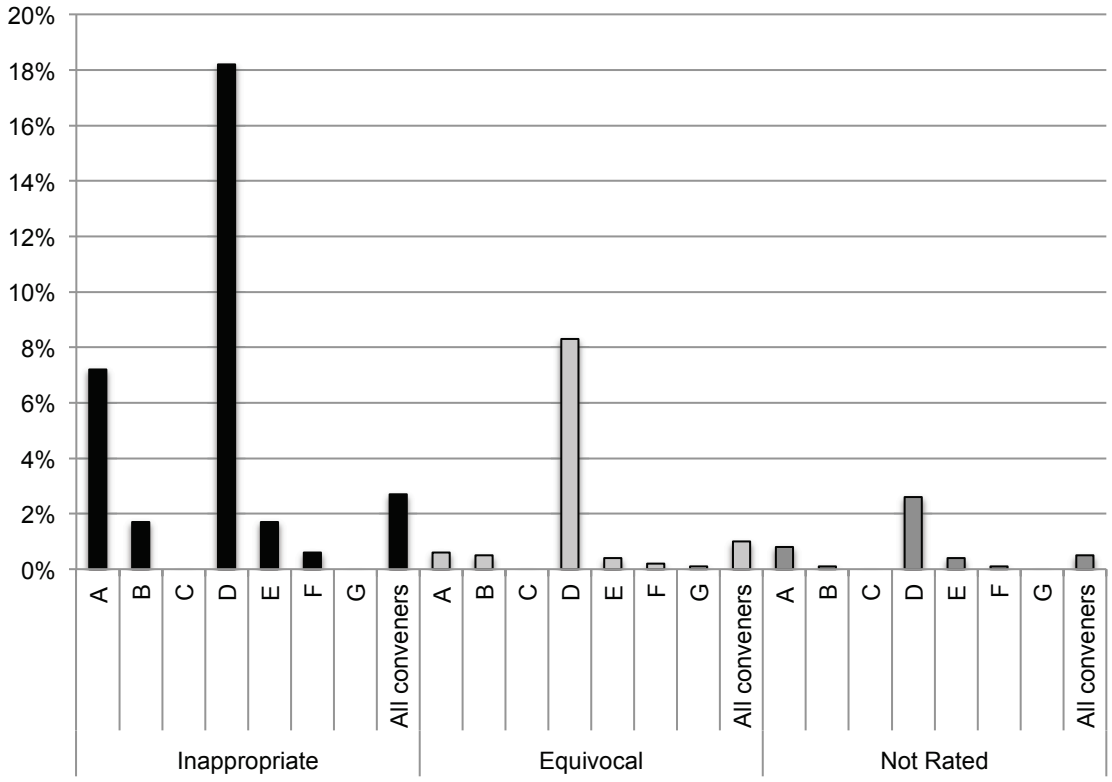
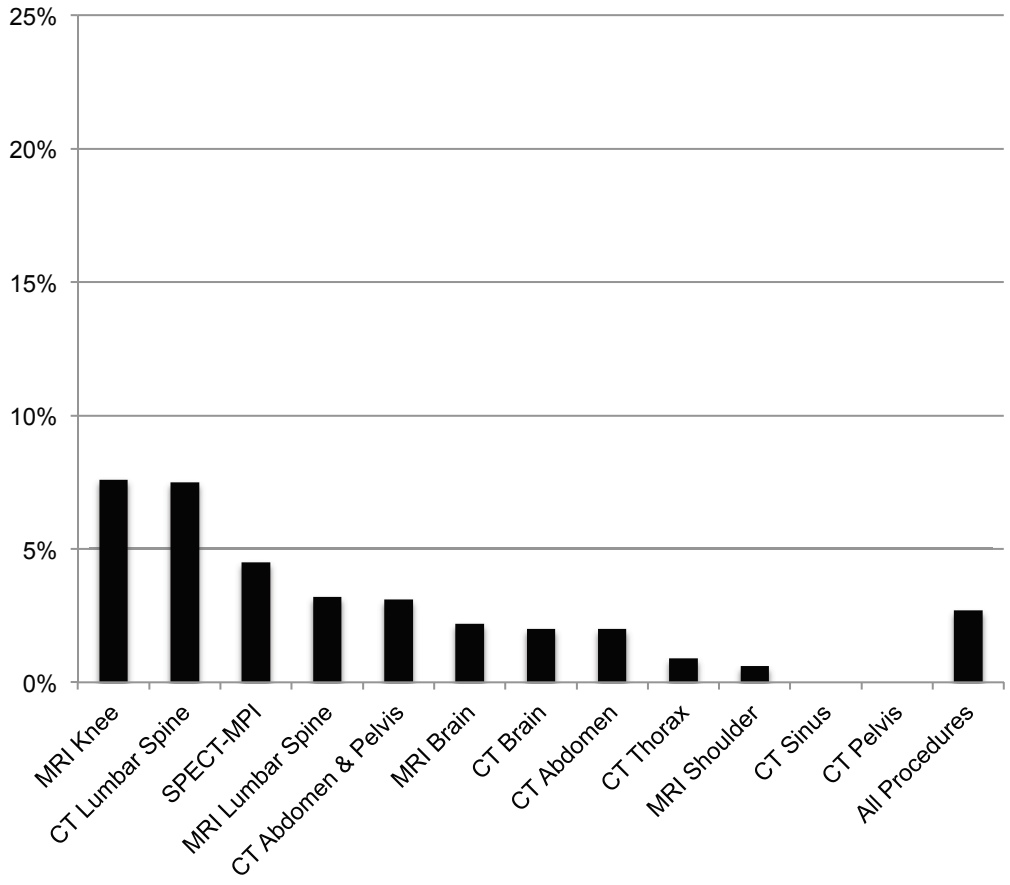


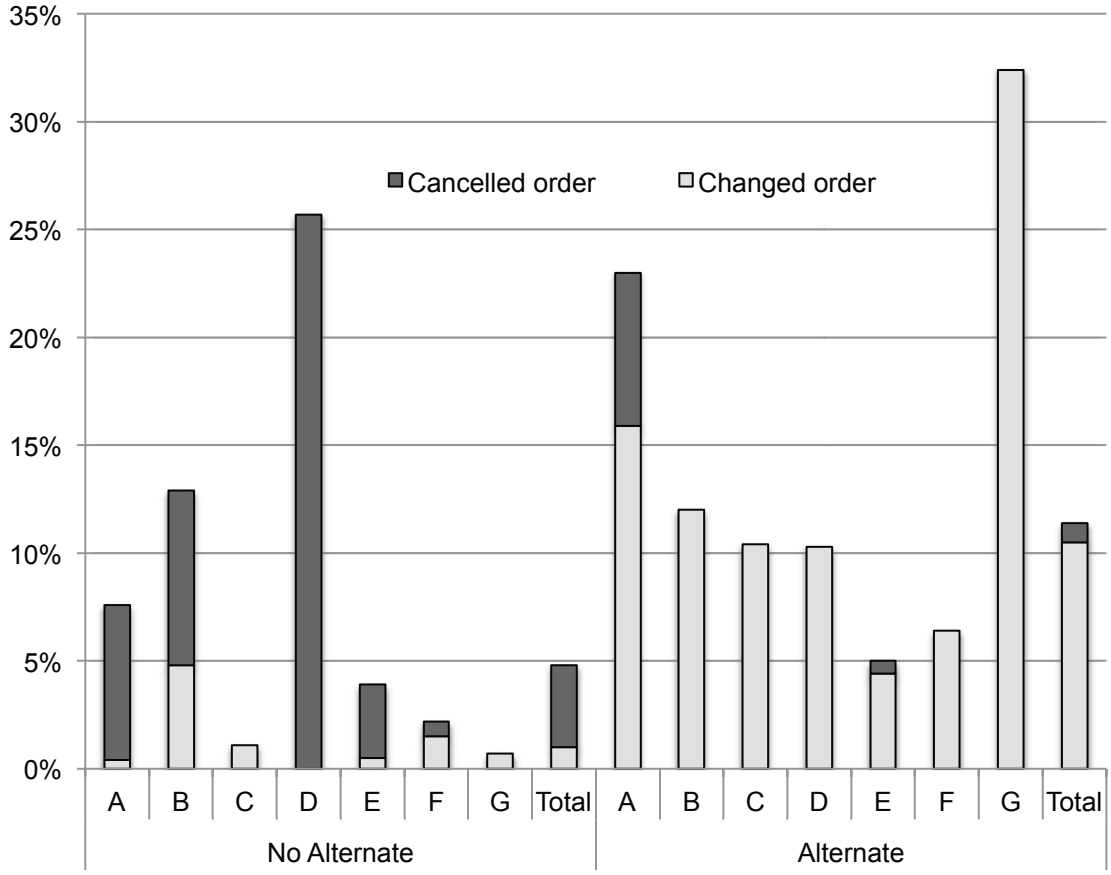
Figure 2.8. Percentage of Inappropriate Orders Canceled After the Initial Order, by Procedure, Intervention Period



NOTE: "Initial" orders placed during the demonstration period are displayed. All conveners are included.

A key finding from these analyses was that a clinician's decision to change an order depended highly on the availability of an alternative procedure with evidence of superior appropriateness (Figure 2.9). Among initial orders rated inappropriate, clinicians were much more likely to change their orders if DSSs provided a recommendation for an alternative procedure. For example, clinicians affiliated with Convener G who received initial scores of inappropriate changed their orders 32 percent of the time if an alternative was recommended to them but less than 1 percent of the time when no alternative was available. While cancellations of initial orders were rare, as noted above, clinicians were far more likely to cancel their order when an alternative was not available compared with orders for which DSSs provided an alternative. In fact, clinicians affiliated with Conveners B, D, and F only canceled orders when alternatives were not available; when alternatives were available, clinicians affiliated with these conveners never canceled their order.

Figure 2.9. Cancellations and Changes Following Feedback of an Inappropriate Rating, by Presence/Absence of an Alternative Order



NOTE: The figure displays the percentage of orders that were changed separately for orders in which the DSS recommended an alternative procedure to the ordering clinician and orders for which the DSS recommended no alternative. Only orders placed during the intervention period that received an inappropriate initial rating are displayed.

Clinicians ultimately changed or canceled approximately 200 of nearly 2,900 orders initially receiving inappropriate ratings during the intervention period. Clinicians that received recommendations for alternative procedures and who changed their order switched to a more appropriate order 63 percent of the time. In the remaining cases, the changed order triggered the same (inappropriate) rating (because the clinician ignored the alternative presented by the DSS; 31 percent) or changed to an order for which the DSS could not rate the appropriateness of the procedure (6 percent). The low volume of inappropriate orders undergoing changes or cancellation precludes us from conducting stratified analyses by convener, procedure, or clinician specialty.

Although we might have expected clinicians to modify their orders at higher rates—particularly when an alternative is available—other factors, including patients' characteristics (including comorbidities) may have influenced clinicians to maintain their original order.

Nevertheless, these patterns indicate that for at least a subset of clinicians affiliated with all conveners, DSSs produced behaviors that are consistent with the demonstration's underlying theoretical model of halting orders that are inappropriate when there are no more appropriate procedures available. Despite the low rate of changes and cancellations to orders initially rated inappropriate or equivocal, these results clearly indicate that the recommendation of an alternative procedure by the DSS is an important mediator of changes in appropriateness.

Table 2.8. Disposition of Orders That Were Either Changed or Canceled by Ordering Clinicians, by Initial Appropriateness Rating and Absence/Presence of Alternative Procedures

Change in Order	Initial Order Rated Inappropriate		Initial Order Rated Equivocal	
	No Alternative Recommended by DSS	Alternative Recommended by DSS	No Alternative Recommended by DSS	Alternative Recommended by DSS
Change to Appropriate	2 (2.4)	50 (39.1)	1 (1.6)	228 (44.4)
Change to Inappropriate*	9 (10.6)	31 (24.2)	1 (1.6)	1 (0.2)
Change to Equivocal*	1 (1.2)	13 (10.2)	7 (11.1)	244 (47.6)
Change to Not Rated	3 (3.5)	6 (4.7)	9 (14.3)	12 (2.3)
Cancel	68 (80)	10 (7.8)	43 (68.3)	8 (1.6)
Change to non-MID Procedure	2 (2.4)	18 (14.1)	2 (3.2)	20 (3.9)
All orders	**85 (100)**	**128 (100)**	**63 (100)**	**513 (100)**

NOTE: For initial orders rated inappropriate, a change to an inappropriate order implies that the clinician ordered another inappropriate procedure; the DSS never offered alternatives that had inappropriate ratings. For initial orders rated equivocal, a change to an equivocal order implies that the clinician selected another procedure recommended by the DSS that had an equivocal rating or else ordered a similarly equivocal procedure that was not among the alternative procedures recommended by the DSS.

2.6. Unadjusted Results—Appropriateness

Figure 2.10 (Panels A–G) displays monthly trends in the appropriateness of imaging procedures over the 24-month demonstration period. Each panel of the figure displays, for each convener, the percentage of orders rated appropriate, equivocal, inappropriate, or not rated. These results reflect the appropriateness of the "final" order as opposed to the "initial" order. Thus, in the intervention period, these results reflect any changes made to the initial order by the ordering clinician after receiving feedback. If the demonstration was working as planned, we would expect the percentage of orders that are not rated to decrease over time, and the percentage of appropriate orders to increase during the intervention period.

We observed three main patterns across the seven conveners in appropriateness trends over the 24-month demonstration. First, five conveners (A, B, D, F, and G), appear to have relatively stable trends over both the baseline and intervention periods. The panel displaying trends for Convener D highlights the extremely high rate of orders that are not rated in that convener—a rate that far exceeds other conveners and remained consistent over the entire 24-month period.

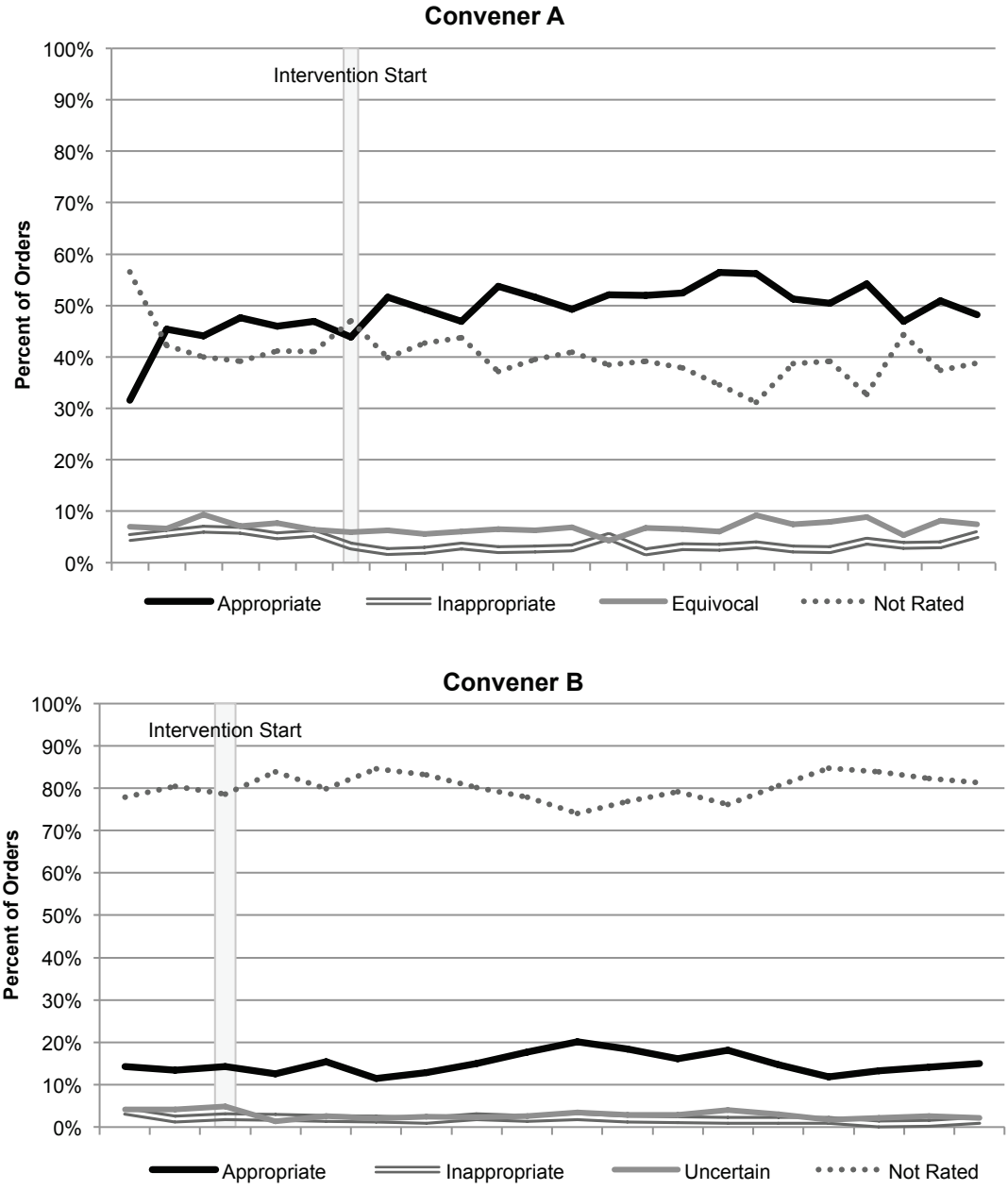
For each of these five conveners, the baseline rate of "inappropriate" and "equivocal" orders was extremely small—suggesting that analyses designed to measure changes in the appropriateness of orders before and after DSSs began providing feedback may have limited power to detect statistically significant differences within each of these conveners.

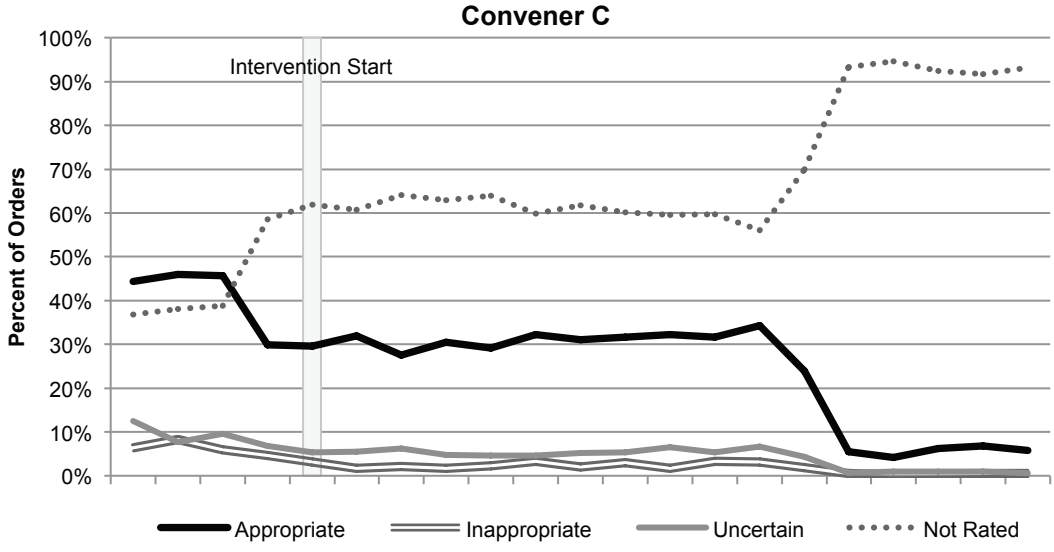
For one convener, we observed an unusual pattern in appropriateness ratings that suggests challenges with the implementation of DSSs. For Convener C, appropriateness ratings followed relatively stable trends for most of the intervention period. However, beginning in month 17 of the demonstration, the percentage of unrated orders increased sharply over a two-month period and then remained at a level of 90 percent or above for the remainder of the demonstration as the use of the DSS was significantly scaled back, as described in Section 2.4.

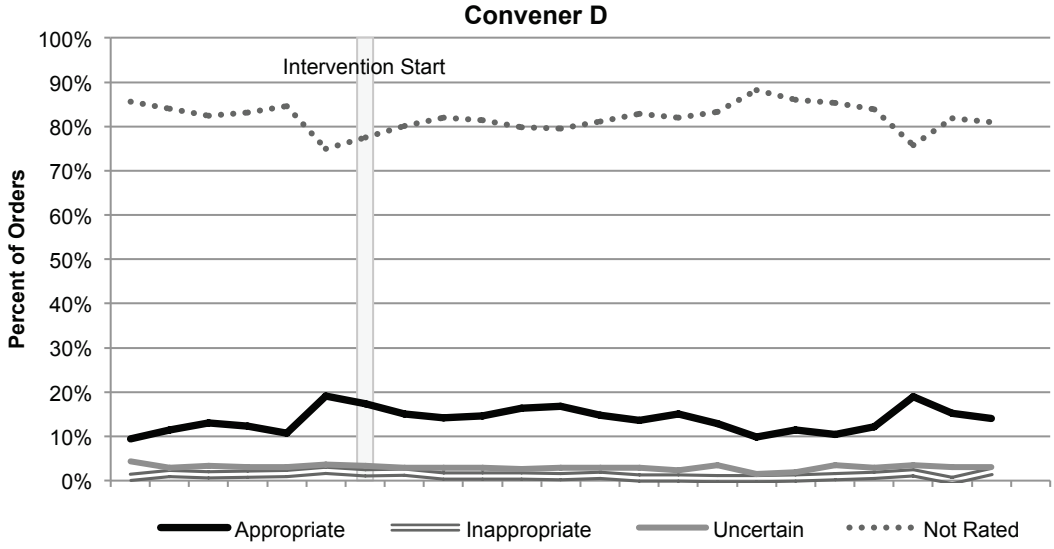
Convener E demonstrates an entirely unique pattern. Although appropriateness ratings for this convener vary from month to month within the first year of the demonstration, the trends stabilize remarkably after the 14th month of the demonstration. From this period onward, the percentage of orders that are rated by the DSS increases by 20 percentage points—the majority of which appear to be orders rated "appropriate." One possible explanation for this pattern is that clinicians affiliated with this convener took longer than expected to master order entry. Although Convener E had the highest rates of inappropriate ordering in the baseline period, the time series plot for Convener E indicates a decline in the percentage of orders rated "inappropriate" over the course of the demonstration.

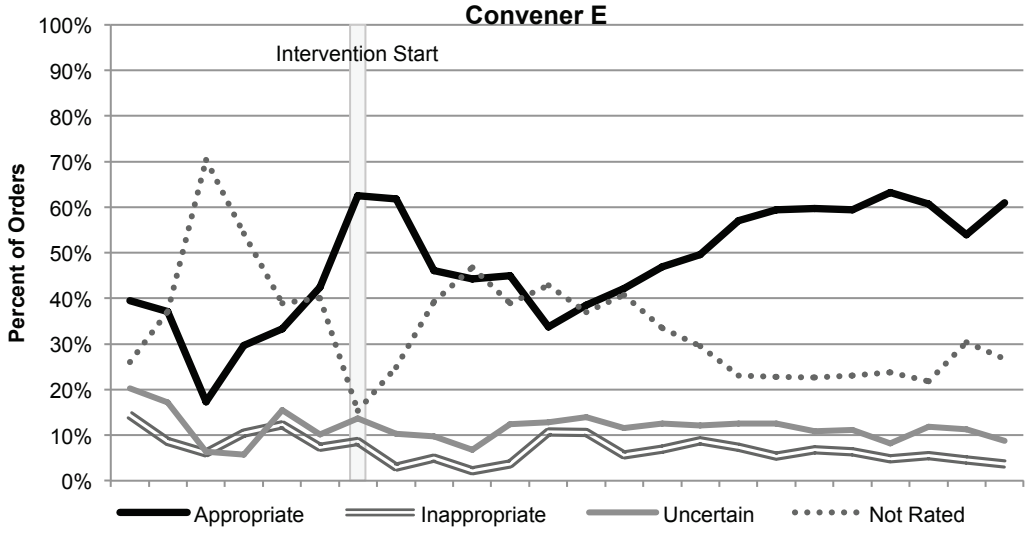
In summary, we found no immediate increases in the rates of appropriate orders or decreases in the rates of inappropriate orders in the intervention period. Moreover, except for the cases noted above, we found relatively flat trends in appropriateness scores over the 18-month intervention period.

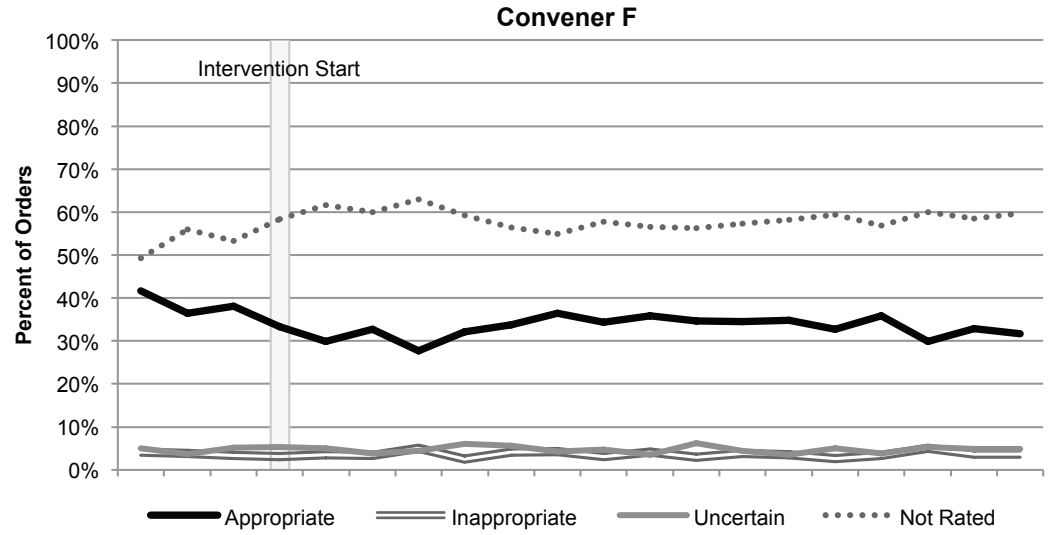
Figure 2.10. Monthly Trends in the Percentage of Appropriate, Equivocal, Inappropriate, and Unrated Orders, Conveners A–G

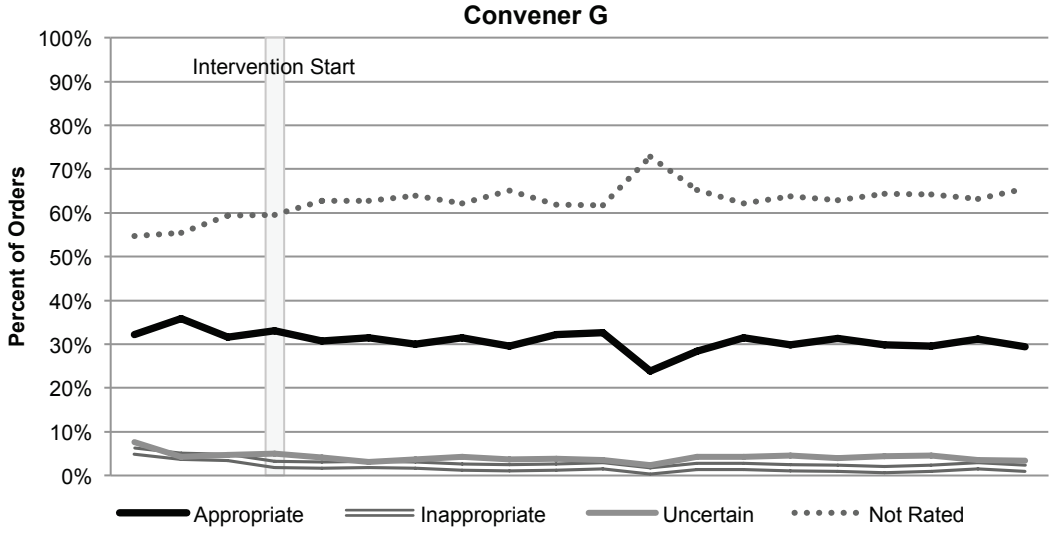



Convener C
Intervention Start
Percent of Orders
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Appropriate
Inappropriate
Uncertain
Not Rated


Convener D
Intervention Start
Percent of Orders
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Appropriate
Inappropriate
Uncertain
Not Rated


Convener E
Intervention Start
Percent of Orders
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Appropriate
Inappropriate
Uncertain
Not Rated


Convener F
Intervention Start
Percent of Orders
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Appropriate
Inappropriate
Uncertain
Not Rated



NOTE: Each panel displays monthly percentages of orders receiving each appropriateness rating (or not rated). The vertical grey line represents the date on which the DSS began providing feedback on the appropriateness of each order (or notified the clinician that the order could not be rated).

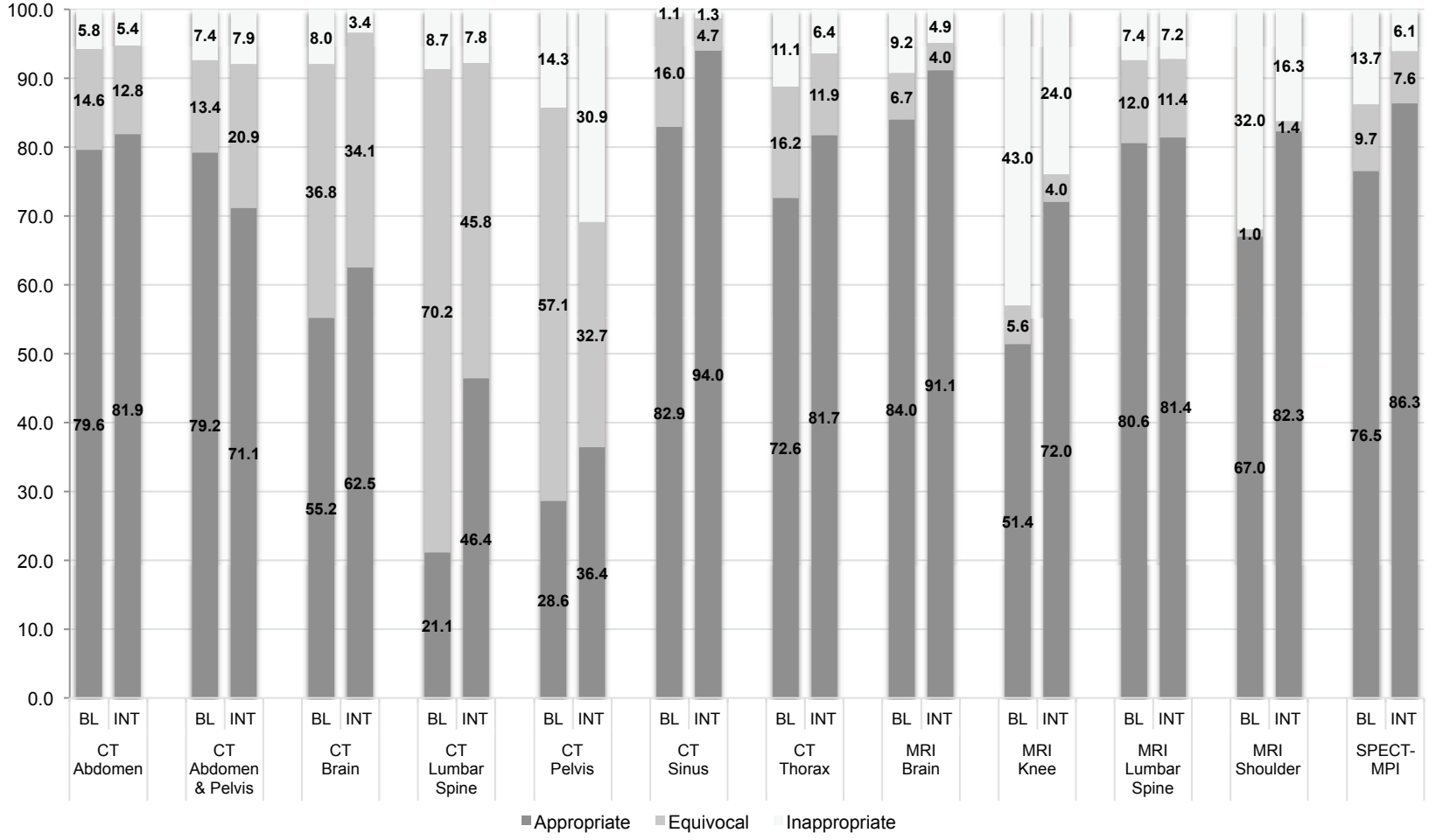
In Tables 2.11 and 2.12, we summarize the changes in the prevalence of orders receiving each "appropriateness rating" between the demonstration and intervention periods. These analyses include only orders for which the DSS was able to match an order to a guideline and assign an appropriateness rating. These results use the "final" appropriateness rating associated with the DSS order, which may reflect changes to the initial order in response to feedback from the DSS. Orders that are canceled are excluded from these calculations. We first present results stratified by imaging procedure and then stratified by convener and clinician specialty.

Overall, among the roughly 8,300 rated orders in the baseline period, 73.7 percent were rated "appropriate." During the intervention period, more than 40,000 orders were rated and 80.7 percent received an "appropriate" rating. The overall improvement (7.0 percentage points) differed widely across procedures. The biggest improvements in appropriateness were for orders for CT lumbar spine and MRI knee (24 percentage points and 21 percentage points, respectively). The only procedure for which appropriateness decreased over time was CT abdomen/pelvis (by 8 percentage points). The procedures associated with the largest decreases in inappropriate ratings were MRI knee (19 percentage points) and MRI shoulder (15.7 percentage points).

These results should be interpreted with caution because most procedures were *less* likely to be rated during the intervention period compared with the baseline period. Thus, these results should not be interpreted as showing that DSS succeeded in changing 7 percent of orders that were "inappropriate" or "uncertain" in the baseline period to "appropriate" orders in the intervention period. Such an interpretation would be valid only if the percentage of orders that were rated remained stable over both periods.

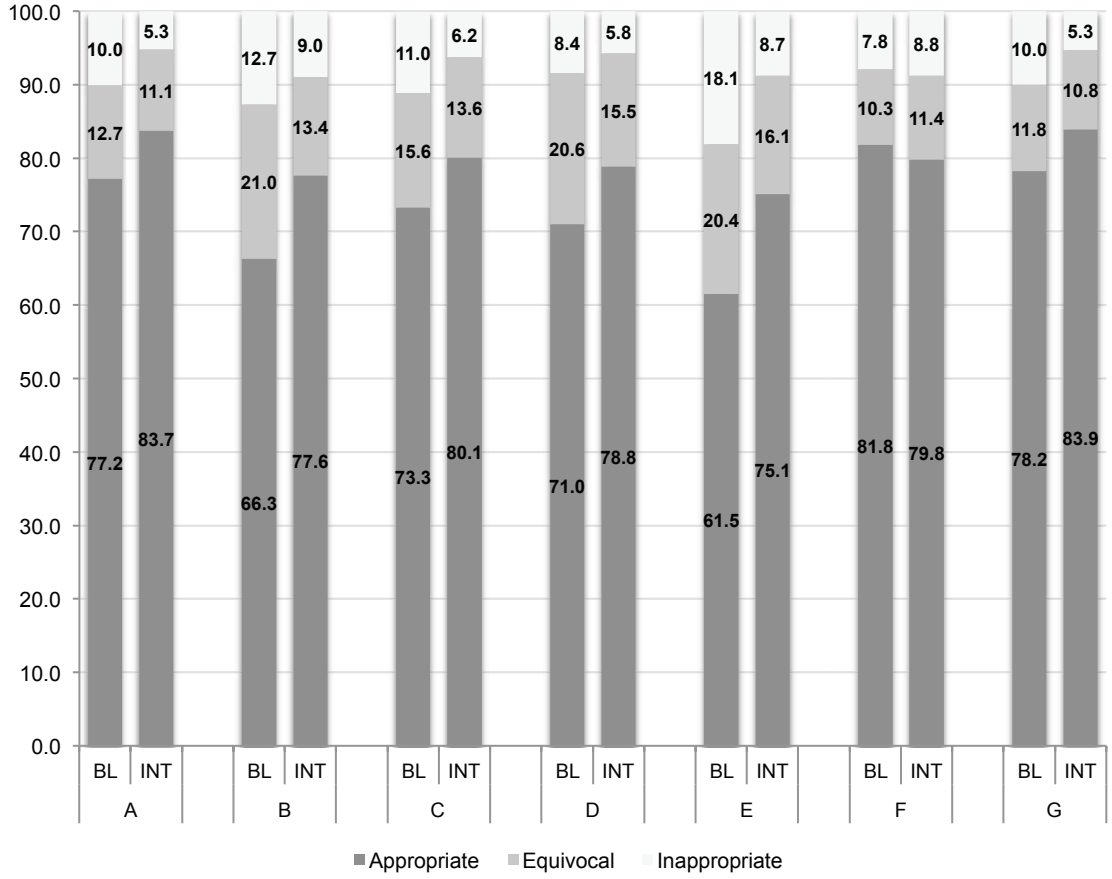
Examining changes in appropriateness ratings across conveners, we found fairly consistent patterns. The percentage of appropriate orders increased for six of seven conveners with magnitudes ranging from 5.7 to 13.6 points between the baseline and intervention periods. Appropriate orders decreased by two percentage points in Convener F. It should be noted, however, that Convener F had the highest rate of appropriate imaging during the baseline period. Convener E was associated with the largest reduction in inappropriate orders (9.4 percentage points) between the baseline and intervention periods.

Figure 2.11. Percentage of Final Orders Rated Appropriate, Equivocal, and Inappropriate, by Demonstration Period and Procedure



NOTE: These results reflect the appropriateness of each clinician's "final" order that may have been changed after receiving feedback from DSS. The denominator for each percentage estimate is limited to rated orders only. BL denotes baseline period; INT denotes intervention period. Among the 8,326 orders placed during the BL, 73.7 percent were rated appropriate, and among the 40,859 orders placed during the INT, 80.7 percent were rated appropriate—a gain of 7.0 percent.

Figure 2.12. Percentage of Final Orders Rated Appropriate, Equivocal, and Inappropriate, by Demonstration Period and Convener



NOTE: Letters A through G denote conveners. These results reflect the appropriateness of each clinician's "final" order that may have been changed after receiving feedback from DSS. The denominator for each percentage estimate is limited to rated orders only. BL denotes baseline period. INT denotes intervention period.

Table 2.10. Percentage of Final Orders Rated Appropriate, by Demonstration Period and Convener

	Baseline Period		Intervention Period		
Convener	Number of Rated Orders	Percentage Rated "Appropriate"	Number of Rated Orders	Percentage Rated "Appropriate"	Change in Percentage Appropriate (Intervention-Baseline)
A	1,515	77.2	5,891	83.7	6.5
B	623	66.3	2,970	77.6	11.3
C	1,943	73.3	6,401	80.1	6.8
D	907	71.0	2,792	78.8	7.8
E	951	61.5	5,773	75.1	13.6
F	920	81.8	5,517	79.8	-2.0
G	1,467	78.2	11,515	83.9	5.7
Overall	**8,326**	**73.7**	**40,859**	**80.7**	**7.0**

NOTE: These results reflect the appropriateness of each clinician's "final" order that may have been changed after receiving feedback from DSS. The denominator for each percentage estimate is limited to rated orders only.

To better understand whether changes in appropriate ordering were broad-based or confined to certain types of clinicians, we further stratified our convener-level analyses by specialty (Table 2.11). Notably, the reduction in appropriate orders we observed within Convener F occurred within most specialty categories with the exception of nonphysician specialists. Among all other conveners, we found fairly consistent improvements in rates of appropriate ordering between baseline and intervention periods for both physicians and nonphysicians and across generalists and subspecialists with only a few exceptions. For example, nonphysicians generalists in Convener A had an 11-point reduction in appropriate orders between the baseline and intervention periods, although the sample of nonphysician generalists in conveners A and F were quite small. An unexpected finding was that medical subspecialists improved their rate of appropriate ordering as much as or beyond that of their generalist counterparts in several conveners. We assumed that generalist physicians would be the most receptive to DSSs and would therefore experience the largest increases in appropriate ordering.

Table 2.11. Change in Percentage of Final Orders Between Baseline and Intervention Periods That Are Rated Appropriate, by Convener and Specialty

	Physicians			Nonphysicians		
Convener	Generalist	Medical Specialist	Surgical Specialist	Generalist	Medical or Surgical Specialist	Overall
A	9.6	8.3	-2.0	-10.7	10.4	6.5
B	9.8	17.5	6.2	7.5	2.5	11.3
C	7.6	11.8	10.6	0.6	3.0	6.8
D	12.4	4.9	5.1			7.8
E	11.8	25.4				13.6
F	-3.0	-0.8	-4.0	-12.0	15.5	-2.0
G	8.9	5.4	3.3	7.4	4.2	5.7
Overall	**8.2**	**9.0**	**5.3**	**1.5**	**5.3**	**7.0**

NOTE: A minimum of 100 orders was required for this analysis. For each cell, the denominator is the total number of scored images for each convener/specialty combination.

Our final set of unadjusted analyses explored the degree to which a clinician's level of exposure to DSSs might diminish or strengthen the observed effect of DSSs on rates of appropriate ordering (see Table 2.12). We divided clinicians into three categories based on their ordering volume and then compared changes in the percentage of appropriate orders between the three groups over the course of the demonstration. We hypothesized that clinicians who ordered less than 20 advanced imaging procedures over the 24-month study period may not have been adequately exposed to the intervention and may exhibit little or no change in their ordering patterns. Meanwhile, clinicians who placed 50 or more orders in a DSS might be more likely to have developed a clear opinion about the value of the feedback. However, we did not observe a consistent pattern of increases in the appropriateness of imaging across the three volume

categories. In Conveners A and B, greater exposure to a DSS is associated with larger changes in appropriateness, but for Conveners C, D, and G, we observed little relationship between a clinician's ordering volume and changes in their rate of appropriate ordering.

Table 2.12 Change in Percentage of Final Orders that are Rated Appropriate, by Clinician-Level Order Volume and Convener

	Number of Orders Placed Through DSS		
Order Placed by	**<20 Orders**	**20–50 orders**	**≥ 50 Orders**
Number of clinicians	3,427 (66.8)	932 (18.2)	769 (15.0)
Number of orders	7,455 (15.2)	11,561 (23.5)	30,169 (61.3)
Convener			
A	3.6	7.2	7.2
B	5.4	10.3	15.8
C	5.9	8.1	6.2
D	8.2	6.6	7.0
E	13.3	21.4	11.2
F	-0.3	-9.4	-0.2
G	6.2	5.8	5.6
Overall	**5.6**	**7.5**	**6.8**

NOTE: Cell values represent changes in the percentage of orders rated "appropriate" between the baseline and demonstration periods.

2.7. Conclusion

These descriptive analyses highlight the diversity of clinicians and practices participating in the demonstration overall and across the seven conveners. Understanding these differences, along with the unique design of each convener's DSS system, revealed to us systematic differences that were independent of DSSs, and that could confound our ability to detect an effect of a DSS on appropriateness and volume if one existed. The more we learned about the differences between conveners, practices, and providers, the more we recognized that stratified analyses defined by convener and specialty would enhance our ability to define meaningful patterns compared with a pooled analysis that could mask effects because of differing directions or magnitudes across subgroups. For example, if the effect of a DSS differs in direction or magnitude between primary care physicians and specialists, the different proportions of specialty types that order MID images across conveners could obscure important findings. This is especially true because of the endogeneity between the type and volume of orders placed, specialty type, use of physicians or nonphysicians to place orders, type of DSS, use of electronic health records integrated with or distinct from DSSs, and the degree of provider engagement with MID. Our analyses showed that specialists typically ordered substantially larger volumes of studies through DSSs while generalists typically ordered a broader set of procedures. These findings suggested that we might expect to see different effects of DSSs across different types of

providers. Our analyses confirmed this hypothesis, further supporting our decision to conduct convener-specific analyses stratified where possible by specialty type.

The extremely high rate with which DSSs were unable to associate an order with a guideline and assign an appropriateness rating significantly complicates our ability to estimate the impact of DSSs. First, it lowers our statistical power to detect an effect of the intervention because we are limited to one-third of our sample of DSS orders (i.e., rated orders) for many analyses. Second, if orders that were rated "inappropriate" or "equivocal" during the baseline period were more likely to be unrated in the intervention period for reasons that are related to the intervention itself (such as greater use of shortcuts by clinicians to bypass the DSS), our conclusions about changes in appropriateness may be biased by limiting our focus to rated orders only.

We examined clinicians' decisionmaking in response to feedback on appropriateness and found that most clinicians whose orders were rated "inappropriate" or "equivocal" continued with their initial orders without any changes. In many cases, clinicians disregarded both the appropriateness rating and one or more recommendations of alternative procedures that had equivalent or higher appropriateness ratings. Nevertheless, a small percentage of clinicians changed their initial inappropriate order when DSSs presented alternatives, and canceled their order when alternatives were not available. These clinicians illustrate that a DSS can work as it was intended. In analyses planned for the near future using clinically rich data that RAND recently collected from all conveners, we will explore patterns in clinicians' decisionmaking within more homogeneous groups of clinical conditions. These "case studies" will help eliminate much of the remaining heterogeneity that might be masking a potentially larger impact of the intervention.

Analyses of trends in the appropriateness of orders for advanced imaging, which was limited to orders that received appropriateness ratings, showed improvements over time of about 7 percentage points. Six of seven conveners achieved higher rates of appropriate ordering, while the seventh convener already had the highest rate of appropriate orders at baseline (82 percent). These increases in appropriateness are indicative of a successful intervention, to the extent that the proportion of orders that are successfully rated by a DSS remains stable between periods. These caveats highlight the need to explore changes in appropriateness within specific clinical conditions to better understand how changes in the percentage of rated orders impacts our ability to draw inferences about the impact of DSSs.

Despite the inclusion of orders from more than 5,000 clinicians, nearly two-thirds of clinicians in our sample placed fewer than 20 orders—suggesting that many clinicians might not have been adequately exposed to the intervention to influence their ordering behavior. However, when comparing clinicians with higher and lower levels of exposure, we found no evidence of a dose-response effect.

This page is intentionally blank.

3. Analysis of the Impact of the MID Demonstration on the Appropriateness of Advanced Imaging Orders

This chapter addresses the research question:

- Were any patterns or trends evident in the appropriateness or inappropriateness of advanced imaging procedure orders?

This component of the analysis of demonstration DSS data evaluates changes with time in the distribution of advanced imaging orders being rated, as compared with not being rated, by (linked to) national specialty guidelines—and, if rated, the distribution of advanced imaging orders across three appropriateness categories (appropriate, equivocal, and inappropriate). We first build models to quantify changes in clinician ordering behavior with respect to the *rated/not rated* dichotomy over the study period, and separately with models of the appropriateness ratings themselves stratified by convener and provider specialty. This chapter complements the previous chapter by introducing regression-based models to account for heterogeneity in ordering patterns.

In the following discussion, we describe models that estimate changes in convener-level aggregate appropriateness levels after the implementation of DSS feedback relative to the initial six-month baseline demonstration period where a DSS does not give providers information on the appropriateness of their orders. These models will be stratified in several ways but chiefly by convener and typically by the provider's specialty.

Modeling strategies must be able to address the very substantial fraction—typically more than one-half—of orders that are not given an appropriateness rating. Hence, we begin by modeling the probabilities of orders being rated at all, using models that are analogous to the model of appropriateness itself. While we would like to know how the appropriateness ratings would change after DSS feedback is provided across all orders—not only those that are given an appropriateness rating—we do not have data to support such an analysis. We will therefore report changes in appropriateness and the fraction of orders that are rated separately, as well as in combination, such that there are four possible outcomes for any order: Unrated, Inappropriate, Equivocal, or Appropriate.

3.1. Unrated Orders

Rating orders for appropriateness is a two-step process. In order for an appropriateness value to be given, the information collected by a DSS must first be linked to guidelines against which the appropriateness can be determined. However, in many cases, the DSS cannot map the order to guidelines. Such orders are therefore unrated.

We first build models to quantify changes in clinician ordering behavior with respect to the rated/not rated dichotomy over the study period, with orders as the unit of analysis. The model itself is a logistic regression model controlling for the baseline period. In addition to the demonstration period and stratifying variables, the model includes random effects for clinicians, which account for heterogeneity in ordering patterns for clinicians within a single specialty.

Table 3.1 summarizes the results of the rated/not rated models. Notably, there is substantial heterogeneity among the providers. Conveners C and G see significant drops in the probability of orders being rated (indicated by odds ratios below 1.0), and this negative change is largely consistent across specialties within these conveners. On the other hand, conveners A, D, and E see changes toward relatively more rated images, and this change is also rather consistent between specialties within the conveners. The evidence is somewhat more mixed with Conveners B and F. These two conveners see no consistent change overall and are mixed in comparisons between specialties.

To better quantify the changes in probabilities of rated images by convener, we calculate marginal effects for this outcome. Such calculations capture expected overall probabilities (of rated images, in this case) for the baseline and intervention periods, averaging over the distribution of all other covariates. Hence, comparing the marginal effects in the probability scale gives a summary of how the probabilities of rated images would have changed from the baseline to intervention periods, all else being equal.

To maintain the spirit of the models that are stratified by specialty while still producing a single summary for each convener, we control for specialty in addition to the pre-post DSS feedback indicator that is of primary interest. For compatibility with the appropriateness calculations below, we perform these marginal effect calculations with the random effects set at zero; this allows the estimates to be interpreted as the predicted proportion of rated images for the "typical" clinician (i.e., one who is neither more nor less likely to have rated images than his or her within-specialty and convener peers) in the baseline and demonstration periods.

Table 3.1. Odds Ratios in a Model of Scored/Not Scored Images, Stratified by Convener and Specialty or Specialty Group with Random Intercepts by Provider [a]

Conveners:	A	B	C	D	E	F	G	Total order volume (all conveners)[a]
All	1.14*	0.99	0.39*	1.38*	3.16*	0.96	0.74*	131,255
Physician: Generalist (total)	0.95	0.72*	0.36*	1.27*	3.27*	0.84	0.79	27,151
Internal Medicine	0.86	0.95	0.48*	1.27*	3.30*	0.71*	0.84	16,305
Family Medicine	1.03	0.47*	0.30*	1.19	3.27*	—	0.77	8,269
Geriatric Medicine	2.07	0.33*	—	1.33	2.70	0.88	0.34*	1,501
Other primary care	1.48	1.51	0.19*	1.87	—	1.78	1.40	1,076
Medical Specialist (total)	1.48*	0.99	0.44*	1.63*	4.26*	0.99	0.70*	59,854
Oncology	1.50	1.74*	0.58*	0.64	5.50*	1.19	0.63*	25,827
Cardiology	3.99*	0.22*	0.21*	1.76*	—	2.20*	0.75	9,674
Neurology	0.34*	0.86	0.51*	2.45*	4.62*	0.69*	0.80	8,335
Pulmonology	1.68*	0.75	0.29*	1.16	5.36*	1.07	0.85	5,997
Gastroenterology	0.27	2.87	0.44	1.24	1.65	1.74	0.36*	2,345
Other medical specialty	0.69	0.73	0.29*	0.89	4.41*	0.35*	0.76	7,676
Physician: Surgical Specialist (total)	1.04	1.16	0.40*	1.29*	>10	1.08	0.98	20,029
Urology	1.76*	0.66	1.13	1.26	—	1.41	2.33*	4,260
Orthopedic Surgery	0.64	1.33	0.38*	0.91	>10	—	0.64	2,728
Otolaryngology	3.44	1.10	0.45*	1.83	—	—	1.37	2,488
Thoracic Surgery	—	—	0.63	1.15	—	1.71	0.67	1,643
Other surgical specialty	0.79	1.71	0.23*	1.32	—	0.73	0.58*	8,910
Nonphysician: Generalist	0.77	1.44*	0.36*	0.72	6.81	0.75	0.65*	19,076
Nonphysician: Specialist	1.15	1.88*	0.37*	0.31	—	0.48	0.75	5,145
Significant specialty OR>1	3	3	0	3	6	1	1	
Significant OR<1	1	3	13	0	0	3	5	

NOTE: OR stands for odds ratio.
* Significant at the p< 0.05 level of significance.
[s] The total number of orders is the volume used in analyses, summed across all conveners.

From the results in Table 3.2, we see that Convener C sees a rather large drop in the proportion of images that are rated—20 percentage points—and Convener G has a substantial drop of 6.3 points. On the positive side, Convener E's probability of rated images increases by 24.6 percentage points. All of these changes are significant except for Conveners B and F. In an absolute scale, the probability of rated images in both the baseline and intervention periods remains below 50 percent for all conveners except A (58.7 percent baseline and 61.2 percent intervention) and E (72.2 percent intervention).

Table 3.2. Calculated Probability of Rated Images for the Typical Provider, Averaged Across Specialty, in the Baseline and the Intervention Periods and the Change Between These Probabilities, by Convener

Convener	Baseline	Intervention	Change
A	0.587	0.612	0.025*
B	0.164	0.162	–0.001
C	0.454	0.254	–0.201*
D	0.139	0.177	0.038*
E	0.477	0.722	0.246*
F	0.426	0.421	–0.005
G	0.399	0.336	–0.063*

* Significance with P<0.05.

3.2. Models of Appropriateness

We now turn to a model of the appropriateness of the rated images, such that the response variable can take on one of the values *inappropriate, equivocal,* or *appropriate*. Our primary statistical tool for this analysis is an ordered logistic regression model. The ordered model accounts for the fact that appropriateness is a coarsened version of a finer (perhaps continuous) underlying appropriateness scale. The most consequential assumption of the model is that the ordered model assumes that covariates either improve or degrade appropriateness ratings across the entire appropriateness scale and by the same proportion across categories of the scale. For instance, if, with the inclusion of a covariate, the probability of an equivocal rating increases relative to inappropriate, the ordered model implies that the probability of an appropriate rating will increase as well.

Exploratory analyses indicate that there is substantial heterogeneity in DSS ordering behavior by provider specialty. Therefore, many of our models will be stratified on this basis. Within specialties, there is also a substantial amount of heterogeneity from provider to provider, accounted for using provider-level random intercepts. Qualitatively, the model says some providers' orders tend to be more or less appropriate than is typical, but that the treatment effect (in the logistic scale) does not depend on the values of the random effects.

Mathematically, our model is as follows: For image *j* of provider *i*, we define an unobserved quantity $z_{ij} = x_{ij}\beta + \nu_i + \varepsilon_{ij}$. The row vector *x* includes covariates such as whether the provider is receiving DSS feedback at the time of the order, the ν values are the provider random effects, which are assumed to have been drawn from a normal distribution with variance learned from the data, and ε is an error term with a logistic distribution that is assumed to be independent across orders and providers. Finally, there are two unobserved cutpoints that are separate from the model describing *z*, called κ_1 and κ_2, whose values are also learned from the data. If z_{ij} falls below κ_1, we observe an inappropriate rating for that image; if z_{ij} falls between the two

cutpoints, we observe equivocal; and, if z_{ij} falls above κ_2, we observe appropriate. All of these models were fit using Stata 13.[9]

3.3. Appropriateness of Rated Orders

Our primary models stratify based on convener and provider specialty, and control for provider-level heterogeneity via random effects. In Table 3.3, we present the results of our main regression models. Starred entries indicate significant differences in appropriateness from the baseline demonstration period to the intervention demonstration period. The values in the table are estimated odds ratios. The odds are defined to be the ratio of the probability of a rated image's appropriateness score being appropriate, divided by the probability of being equivocal or inappropriate. The odds ratio is then defined to be the odds for the intervention period divided by the odds for the baseline period. The ordered logistic model assumes that the ratio of the probability of appropriate to the probability of equivocal or inappropriate is equal to the ratio of the probability of inappropriate or equivocal to the probability of inappropriate. Hence, the reported odds ratios summarize increases or decreases in appropriateness level: Values greater than 1 indicate an increased probability of higher appropriateness ratings (i.e., appropriate versus inappropriate or equivocal, or appropriate or equivocal versus inappropriate) and odds ratios smaller than 1 indicate increased probability of lower appropriateness ratings among rated orders.

For some convener/specialty combinations, there was some combination of insufficient number of orders, too few providers, or too little heterogeneity in the results so we were unable to fit the model. Such cases are marked by "—." The first row—labeled "All"—includes data from all specialties. The table's rows have been arranged within specialty group by total rated image volume (except for the catch-all "other" categories). There are substantial differences in volume across specialties, with 7,000 rated images for internal medicine (a subset of generalist physicians), but barely more than 400 for gastroenterology (a subset of medical subspecialist physicians).

Several broad features of the results merit mention. First—and most notably—there are only two estimated odds ratio below 1 (indicating worsening appropriateness) that are statistically significant (Convener D's pulmonology and Convener F's "other" medical specialty). Since we are reporting results from many dozens of models, it would not be surprising to have two significant, negative results even if the effects of a DSS were neutral to positive in all situations. On the other hand, 22 of the estimates are statistically significant and positive for the mutually exclusive specialties, not including the larger aggregate specialty groups (for example, medical specialty, or surgical specialty noted with "(total)" in Table 3.3). Moving down the rows of the table within specialty groups, statistically significant results are less frequent, which may be the

[9] Additional information is available at stata.com.

result of decreasing statistical power, rather than a weakening treatment effect. Without controlling or stratifying for specialty, we see statistically significant improvements in the appropriateness rate for all except Convener F.

Table 3.3. Ordered Logistic Regression Results for the Ordered Appropriateness Outcome Among Rated Orders

Conveners:	A	B	C	D	E	F	G	Total # of rated orders[a]
All	1.71*	1.67*	1.63*	1.52*	2.26*	0.91	1.56*	46,705
Physician: Generalist (total)	1.88*	1.57	1.42*	1.73*	1.75*	0.80	1.85*	11,642
Internal Medicine	1.70*	1.61	1.14	1.81*	1.67*	0.73	1.80*	7,098
Family Medicine	2.25*	1.84	1.49*	1.56	1.82*	<.01	3.35	3,661
Geriatric Medicine	2.02	0.55	—	1.44	3.18	0.98	2.67	496
Other primary care	5.91	1.34	>100	<.01	—	0.99	<.01	387
Physician: Medical Specialist (total)	2.04*	2.32	2.17*	1.26	3.10*	1.02	1.61*	21,421
Cardiology	2.16*	1.86	2.87*	1.67	3.47*	2.43	1.66	6,221
Oncology	0.63	1.35	2.81*	0.58	—	0.80	2.29*	4,776
Neurology	—	36.02	2.75	<.01	0.41	0.94	1.31	3,677
Pulmonology	3.70*	11.85	0.95	0.21*	0.33	1.82*	1.19	3,604
Gastroenterology	0.66	2.42	2.01*	0.93	0.72	1.25	1.77	413
Other medical specialty	0.73	1.07	2.37*	2.30*	2.85	0.33*	0.92	2,730
Physician: Surgical Specialist (total)	1.16	1.76	2.53*	1.77*	9.01	0.69	1.81*	6,435
Urology	1.40	1.55	30.08*	6.10*	8.04	NA	2.19*	1,520
Otolaryngology	0.55	1.31	0.43	1.20	—	0.68	0.94	1,244
Orthopedic Surgery	14.25	1.50	1.96	4.10*	—	2.75	1.23	1,208
Thoracic Surgery	—	—	1.81	1.23	—	NA	3.14*	717
Other surgical specialty	1.50	2.39	1.95	1.27	—	0.65	0.98	1,746
Non-physician: Generalist	0.78	0.77	1.35*	2.00	—	0.89	1.21	5,565
Non-physician: Specialist	1.34	1.21	1.31	1.88	<.01	1.88	1.38	1,642

* Significant at the P < 0.05 level of significance.

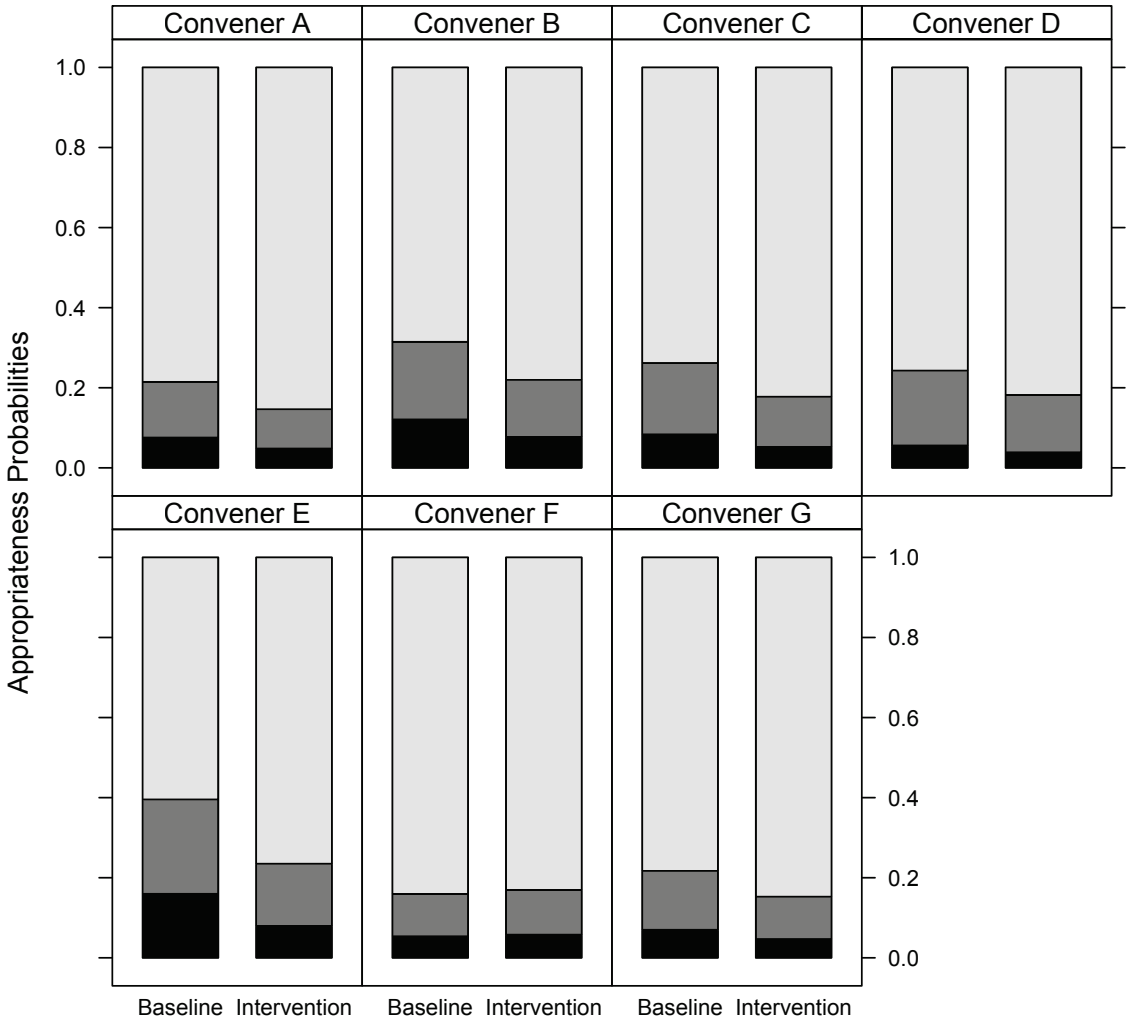
[a] The total number of orders is the volume used in analyses, summed across all conveners.

As a perhaps more easily understood statement of the results, we summarized some of the model results in terms of marginal effects as described above for the rated/not rated analysis. As before, these summaries average over the specialty distribution. The marginal effect summaries are interpreted as the expected change in probabilities for clinicians who are "typical" with respect to their image appropriateness and are consistent with the marginal effects for the rated/not rated model. And, as before, we used a single model per convener controlling for—rather than stratifying by—specialty.

In Figure 3.1, we see these summaries for the typical provider averaged across specialties. Although Convener F does not see an increase in appropriateness for the typical provider, its percent of appropriate images among rated orders compares favorably with all other conveners,

even in the intervention period. The typical provider in all other conveners sees gains in appropriateness. The gain for Convener E is the strongest—jumping from around 60.4 percent to 76.5 percent, though it still lags behind other conveners in the intervention period. In terms of drops in the proportion of inappropriate orders, the change ranges from -0.4 percentage points (Convener F) to 8.0 points (Convener E). These changes in appropriateness (both in terms of probabilities of more appropriate and fewer inappropriate ratings) are significant for all conveners except Convener F.

Figure 3.1. Probabilities of Appropriateness Categories of Rated Orders for Typical Providers, Averaged Across Specialties



NOTE: Light gray is appropriate, dark gray is equivocal, and black is inappropriate. All differences are statistically significant except those for Convener F.

3.4. Combining Rated and Unrated Orders

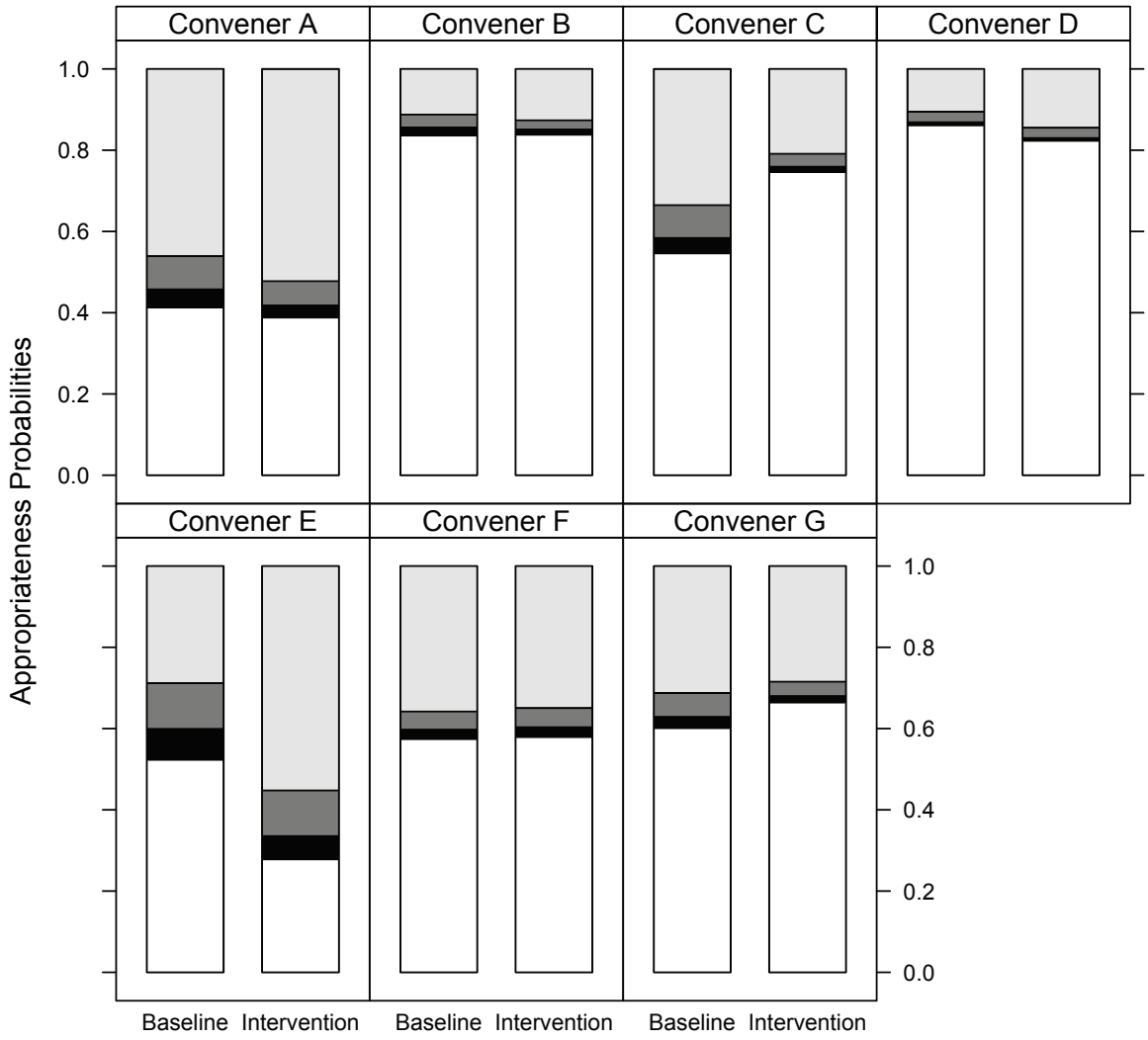
As mentioned above, the evaluation would be substantially strengthened if the appropriateness of all orders—not just the half or fewer that are rated—could be modeled. A primary concern is that even if appropriateness is increasing over time for the rated images, the appropriateness ratings across all images—if they were all somehow rated against prevailing best practices within specialties—may be holding steady or even decreasing across the study period.

While we cannot know how the unrated images would be scored in such an ideal setting, we can look at the aggregate changes in the sense of marginal effects of treatment. In these calculations, the rated/not rated and appropriateness models are assumed to be statistically independent of one another. If that assumption holds, we can make statements about the marginal effects of treatment on the outcome variable whose value is one of not rated, inappropriate, equivocal, or appropriate. Another downside of modeling the rated and appropriateness outcomes separately is that we are not able to assess statistical significance of the marginal differences between the baseline and intervention periods.

Figure 3.2 contains such marginal effects summaries, which are representative of the appropriateness rating distribution of the typical provider, averaged across specialties. Here we can see that even though appropriateness almost uniformly increased from the baseline to intervention periods among the scored images, as a share of all images, the probability of an appropriate rating decreased for typical providers in Conveners C, F, and G. Only for Conveners A, B, D, and E does the typical provider's probability of an appropriate rating increase among all orders in this four-category representation. While it is impossible to know how these numbers would change if all images had been rated, the consistent and favorable changes in terms of the percent of appropriate images among the rated images looks less promising in light of the substantial—and often increasing—proportion of unrated orders.

In summary, these results show that appropriateness levels did improve substantially for rated orders after DSS feedback was turned on. However, a very large proportion of orders were not rated; this fraction increased for two conveners (across all specialties) and within particular specialties for all but two of the conveners. Although these results suggest that providers do adjust advanced imaging ordering practices—or at least adjust their explanations of the reasons for the order—when given DSS feedback, it is not clear how these changes would affect total appropriateness rates of all ordered images.

Figure 3.2. Probabilities for Typical Providers of Combined Rated/Not Rated and Appropriateness Categories, Averaged Across Specialty



NOTE: The calculated probabilities assume statistical independence between the rated/not rated model and the appropriateness of the rated orders. Light gray represents appropriate, dark gray is equivocal, black is inappropriate, and white is not rated.

This page is intentionally blank.

4. Relationships Between the Appropriateness of Advanced Imaging Procedure Orders and Imaging Results

This chapter will discuss the question:

- Is there a relationship between the appropriateness of advanced imaging procedure orders and imaging results?

(Specifically, we examine whether receipt of particular patterns of appropriateness feedback affects the subsequent volume of advanced image orders.)

While CMS and the evaluation team initially hoped the results of advanced imaging studies would be available for analysis, this was not feasible from the perspective of conveners. Without the opportunity to access data describing image test results, we elected to examine whether receipt of patterns of appropriateness feedback affects the subsequent volume of advanced image orders.

To examine whether DSS feedback affects the volume of advanced imaging, we consider whether appropriateness rates early in the intervention period predict utilization in a later period. This chapter supplements the distribution of appropriateness ratings presented in Chapters 2 and 3 by focusing on the specific question of whether exposure to feedback about inappropriately rated orders during the first 90 days of the intervention period for each practice predicts a relative increase or decrease in DSS volume during the next 90 days.

4.1. DSS Feedback Impacts and Volume Measures

In particular, for each clinician, we capture DSS volumes of all orders, rated orders, and appropriate and inappropriate orders. We use rates such as the percent of rated orders that are appropriate in the first 90 days of the intervention period for each practice to predict the relative increase or decrease in DSS volume for the next 90-day period. We measure the relative increase as:

Log(Number of orders in second period/Number of orders in first period).

To stabilize the estimates, we restrict our analyses to providers who have at least 15 orders in the first 90-day period. This includes 284 providers ordering advanced imaging at one of the demonstration sites (or 5.5 percent of the 5,128 providers who appear in our DSS data files to have at least one MID order). Additionally, we drop three providers who met the initial volume requirement but have no advanced image orders in the second 90-day period out of concerns that they may have stopped practicing at a study site and to avoid problems taking the logarithm of zero. Although other providers may have stopped practicing at a study site at some point in the

second 90-day period, we expect such attrition to be independent of ordering behavior and therefore to have minimal impact on our analyses. The thought behind these models is that feedback of inappropriate ratings may discourage inappropriate advanced imaging order in the future.

The focus of this demonstration was to improve the distribution of appropriateness ratings for advanced imaging by reducing the use of inappropriately rated advanced imaging. As a supplement to data about appropriateness ratings reported in Chapters 2 and 3, this chapter uses MID data about variation in providers' early appropriateness portfolios to predict subsequent ordering volume. In this way, Chapter 4 provides an additional measure of the effectiveness of MID's DSS in changing image ordering patterns.

MID provides data showing variation in the degree to which providers receive appropriate, equivocal, and inappropriate ratings for their orders. Using these data, we can construct various definitions of the clinician's experience with DSS-rated orders as shown in Table 4.1 (column 1). This analysis hypothesizes that the pattern of appropriateness feedback that providers receive during the first 90 days of exposure to concurrent appropriateness rating through DSSs affects the volume of subsequent image ordering. For example, row 1 in Table 4.1 examines how the percentage of inappropriately rated orders (i.e., *# inappropriately rated/# rated orders*) affects subsequent ordering volume among all MID providers. Other rows in Table 4.1 highlight other definitions of ordering patterns.

Table 4.1. Impact of DSS Ratings During First 90 Days of Feedback on DSS Volume in Second 90 Days of Feedback

Alternate Definitions of Clinician's Experience with DSS-Rated Orders	Estimate	95% Confidence Interval	Provider Specialty
# Inappropriate orders/# Orders	–0.81	(-2.01, 0.39)	All
# Appropriate orders/# Orders	–0.11	(-0.34, 0.11)	All
# Inappropriate orders/# Rated Orders	–0.16	(-0.45, 0.13)	All
# Appropriate orders/# Rated Orders	–0.06	(-0.26, 0.14)	All
# Inappropriate orders/# Orders	–2.41	(-5.89, 1.07)	Primary care
# Appropriate orders/# Orders	–0.02	(-0.62, 0.56)	Primary care
# Inappropriate orders/# Rated Orders	–0.62	(-1.19, -0.05)	Primary care
# Appropriate orders/# Rated Orders	–0.19	(-0.56, 0.17)	Primary care
# Inappropriate orders/# Orders	–0.53	(-1.79, 0.72)	Specialist
# Appropriate orders/# Orders	–0.13	(-0.38, 0.11)	Specialist
# Inappropriate order /# Rated Orders	–0.02	(-0.36, 0.32)	Specialist
# Appropriate orders/# Rated Orders	0.00	(-0.24, 0.24)	Specialist

NOTE: Alternate Definitions of Clinician's Experience with DSS Ratings serve as covariates for the models in which the clinician's experience with DSS ratings during the first 90 days of their practice's exposure to predicts The "covariate" column describes alternative definitions of an estimated provider-level appropriateness rates. All of the covariate measures are defined using data only from the first 90 days of each practice's exposure to DSS (during the intervention phase of the demonstration).

We hypothesize that those providers who receive higher rates of inappropriate feedback ratings (defined either as a higher proportion of *all ordered* or of *all rated* images being rated as

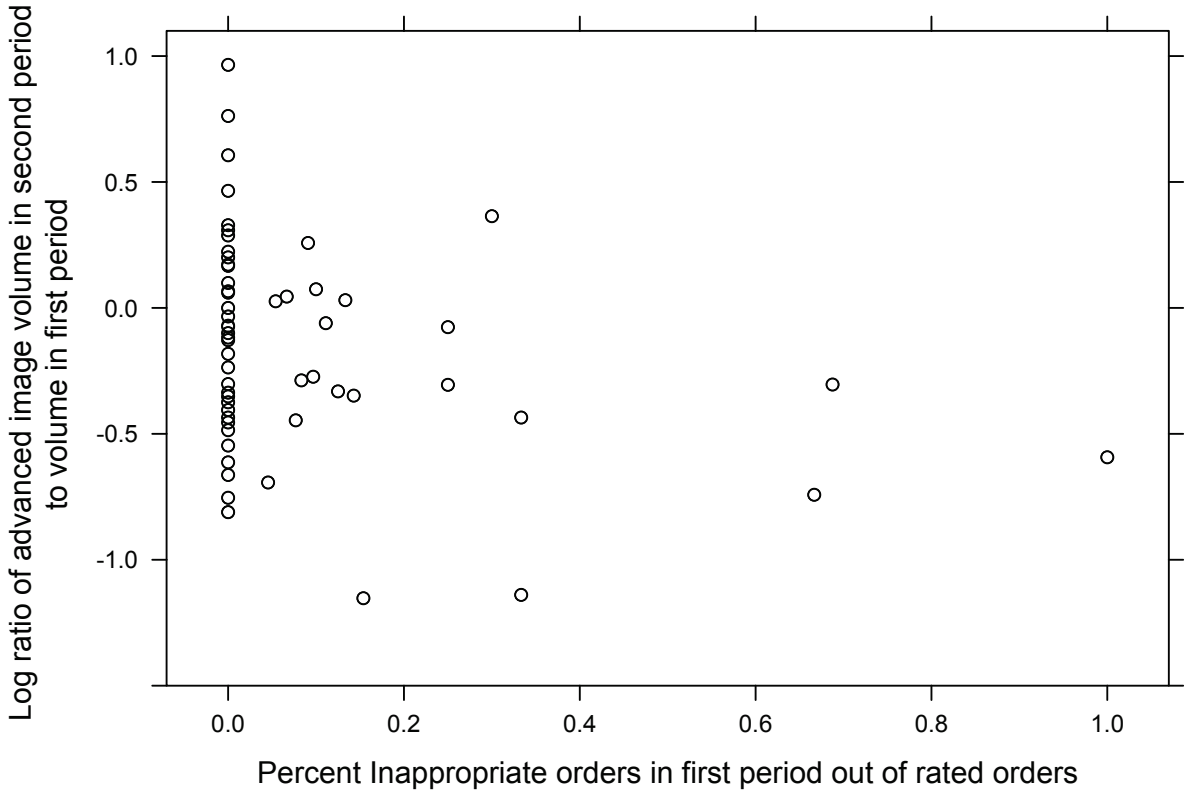
inappropriate), are more likely to drop their future total volume of ordering advanced imaging. Specifically, we hypothesize total volume will drop because clinicians will cease ordering advanced images for reasons that they have learned are inappropriate. This analysis studies whether receipt of higher rates of feedback on inappropriate orders during the first 90 days of exposure to concurrent appropriateness feedback is associated with a drop in total number of advanced images ordered during the subsequent 90 days (i.e., days 91 through 180 of each practice's exposure to appropriateness feedback ratings).

Given that this is a provider-level analysis, we have little ability to stratify the analyses by convener and specialty as we did when analyzing changes in the appropriateness of the orders. We do examine models that stratify on the basis of primary care provider/specialist dichotomy (across all conveners).

Across 12 analyses performed shown in Table 4.1, we observe only one single statistically significant result. Row 7 shows that for primary care providers, we observe a higher proportion of inappropriate orders out of all rated orders in the first 90-day period is weakly associated with a decrease in MID image orders, as measured in a DSS. No other confidence intervals support significant findings. Although it is plausible that primary care physicians' practice patterns would be more strongly affected by DSS feedback than that of specialists, the evidence from the data are quite weak in this regard. Twelve hypothesis tests yielded a single significant result, which would not be anomalous if there truly were no effect of the first period's DSS feedback on the second period's ordering patterns.

Figure 4.1 displays the data from this one nominally significant result. Beyond the fact that there are relatively few providers who meet the volume requirements to be included in the analysis, most who do have no inappropriate orders in that period. Having such a large portion of the analytic sample with identical values of the explanatory variable harms our ability to detect its effect on our outcomes of interest. As with other aspects of the evaluation, our ability to detect such changes would likely be substantially improved if a higher proportion of orders had been rated. In the end, the data provide little evidence that DSS feedback has a substantial impact on individual providers' practice patterns. However, the data do not rule out such impacts, either.

Figure 4.1. Relationship Between the Percentage of Inappropriate Orders in the Initial Period During Which Clinicians Are Exposed to DSS[a] and the Volume of Advanced Imaging Orders[b] in the Subsequent Period[c]



[a] The initial time period during which clinicians are exposed to DSS is defined as the first 90 days during which each practice uses DSS to provide real-time appropriateness rating feedback to those who order advanced images.
[b] Volume of advanced imaging in the subsequent time period is defined as the log ratio of advanced imaging volume during the second time period to volume during the first time period.
[c] The volume of advanced image orders is defined during days 91–180 following each practice beginning to provide real time appropriateness feedback.

As another approach to detecting a favorable impact of DSS feedback on provider ordering practices, we examine the percentage of providers who see improvements in the proportion of inappropriate orders across three time periods: the final 90 days of the baseline period (where no feedback is provided); the first 90 days when DSS feedback is provided; and, the second 90-day period (i.e., days 91 through 180) when DSS feedback is provided. Here, we record whether each provider's proportion of inappropriate orders increases, decreases, or stays the same in each of these periods. (Provider periods with zero DSS volume are dropped from the analysis, so the size of the sample varies by the periods under comparison.)

This analysis is motivated by the small number of providers who have substantial numbers of inappropriate orders. Consider a provider who has 20 advanced image orders in each of two periods, with one inappropriate order in the baseline period and zero during the first 90

days of feedback. Certainly, we cannot attribute the drop in this single provider's inappropriate order rate to DSS feedback. However, if DSS has no impact on providers' ordering habits, we expect that there will be a similar number of providers whose inappropriate rates worsen rather than improve. To be clear, this test only addresses the proportion of providers whose appropriateness rates improve, and does not address overall appropriateness rates, which must account for differences in provider volumes and are considered elsewhere in this report.

In Table 4.2, rows 1 and 4 and 2 and 5, respectively, compare the baseline to the first or second 90 days of the intervention period. These rows show a significant reduction in the percentage of providers whose inappropriateness rates improve rather than worsen or show no change from the baseline period to the first two 90-day periods after DSS feedback was turned on.

Table 4.2. Impact of DSS Feedback on Provider Ordering Practices (A Second Analysis)

		Number of Providers with Inappropriately Rated Orders:			
	Demonstration Period [a]	Lower Percent	Higher percent	Unchanged percent	P-value
Denominator is total DSS volume including both Rated and Not Rated order volume					
1	Baseline Period to Intervention Days 1-90	344	210	1645	< .001
2	Baseline Period to Intervention Days 91-180	331	185	1588	< .001
3	Intervention Days 1-90 to Intervention Days 91-180	230	231	1805	0.96
Denominator is total DSS volume including only Rated order volume					
4	Baseline Period to Intervention Days 1-90	276	160	794	< .001
5	Baseline Period to Intervention Days 91-180	274	149	751	< .001
6	Intervention Days 1-90 to Intervention Days 91-180	192	185	889	0.72

[a] DSS was implemented for a 24-month window so that the DSS generated appropriateness ratings for all orders that linked to guidelines during the entire 24-month window. By design, the demonstration was implemented so that the appropriateness ratings assigned to ordered images during the baseline demonstration period (the first six months of the demonstration) were not known to their ordering clinicians. During the remaining 18 months of the demonstration, appropriateness ratings were provided to ordering clinicians in real time immediately after the order was placed as long as the order was rated (i.e., linked to an existing guideline).

[b] p-values<0.05 indicate significant differences in the percentage of providers whose inappropriate rates are lower rather than higher relative to the expected even-split if DSS feedback did not affect inappropriateness rates.

[c] Results hold regardless of whether the denominator in the inappropriate rate is calculated using all orders or only rated orders.

The finding of a significant reduction in inappropriateness rates even during intervention days 1–90 indicates that any "burn-in" or learning process to adapt to DSS feedback was rather rapid. The finding of a significant reduction in inappropriateness rates for both intervention periods 1–90 days and 91–180 days compared with the baseline period shows the results are sustained. Results hold regardless of whether the denominator in the inappropriateness rate is calculated using all orders or only rated orders.

In contrast, rows 3 and 6, respectively, compare the first 90-day block (intervention days 1–90) with the second 90-day block (intervention days 91–180). DSS feedback was turned on throughout both of these periods. Comparisons presented in rows 1 and 2 respectively show 62 percent and 64 percent of providers reduced the proportion of orders that were assigned an inappropriate rating among providers who saw a change in the percent of their orders that were inappropriately ordered. In contrast, row 3 shows no change in the proportion of orders that were assigned an inappropriate rating among providers who saw a change in the percentage of their orders that were inappropriately ordered. While the treatment effect is not escalating over time, neither is it eroding.

4.2. Conclusion

The DSS volume analysis described in this chapter corroborates the evidence elsewhere in the report that a DSS's impact on appropriateness of ordered images (Chapters 2 and 3) is stronger than its impact on volume (Chapter 5).

Section III: Convener-Level Analyses of Advanced Image Utilization Before and After the Medicare Imaging Demonstration Was Introduced in Practices Associated with the Demonstration Compared with Comparable Control Practices

This section addresses the statute question:

- Were any national or regional patterns or trends evident in utilization of advanced imaging procedures? (Chapter 5)

This page is intentionally blank.

5. Trends in Imaging Utilization in the Medicare Imaging Demonstration

This chapter discusses the statute question:

- Were any national or regional patterns or trends evident in utilization of advanced imaging procedures? (Chapter 5)

To examine national and regional trends in advanced imaging utilization, we compared trends in imaging utilization over time between demonstration group ordering clinicians and a matched set of control ordering clinicians who were not exposed to DSS (see Appendix A). We measured utilization between January 2009 and November 2013, a period that includes the predemonstration period and the demonstration baseline and intervention periods. The predemonstration period represents 21 months (January 2009 through September 2011) prior to the demonstration's initiation. The six-month baseline period, October 2011 to April 2012, was followed by an 18-month intervention period.

In this chapter, subsection 5.1 presents trends in utilization in the demonstration group, stratified by several important dimensions. Subsection 5.2 presents results of statistical tests for a demonstration effect on trends in advanced imaging utilization. Subsection 5.3 presents a brief discussion of the findings presented in this chapter.

5.1. Utilization of Advanced Imaging as Measured Using Claims

We measured utilization at the ordering clinician level. In each month, we counted the number of imaging procedures ordered by each ordering clinician in the sample, as indicated by the referring provider field and date of service on carrier claims.

We compared trends in utilization between a "demonstration group" and a "comparison group" of ordering clinicians. The demonstration group included 5,744 clinicians identified as associated with practices participating in the demonstration using the tax identification number on claims and who ordered MID imaging procedures in the 18 months prior to the demonstration. The comparison group included 41,538 unique clinicians selected from counties matched to demonstration counties and who ordered MID imaging procedures in the 18 months prior to the demonstration. Appendix A provides more detail on the identification of demonstration and comparison group ordering clinicians.

MID imaging procedures were defined on the basis of professional component and global service claims for defined procedure codes, excluding procedures performed in an emergency department or inpatient hospital (see Appendix A). To further examine broader imaging utilization patterns, we also created measures of utilization of "non-MID imaging procedures,"

defined as imaging procedures not included in MID and identified as substitutes for MID imaging procedures by two primary care physicians on our team (Kahn, Wenger) and a radiologist consulting with our team (Lee). We calculated rates of utilization of MID and non-MID imaging procedures per unique Medicare beneficiary treated per ordering clinician per month. The denominator of this measure is the number of unique Medicare beneficiaries treated per physician per month (a beneficiary is identified as "treated" if they received a procedure or evaluation and management visit).

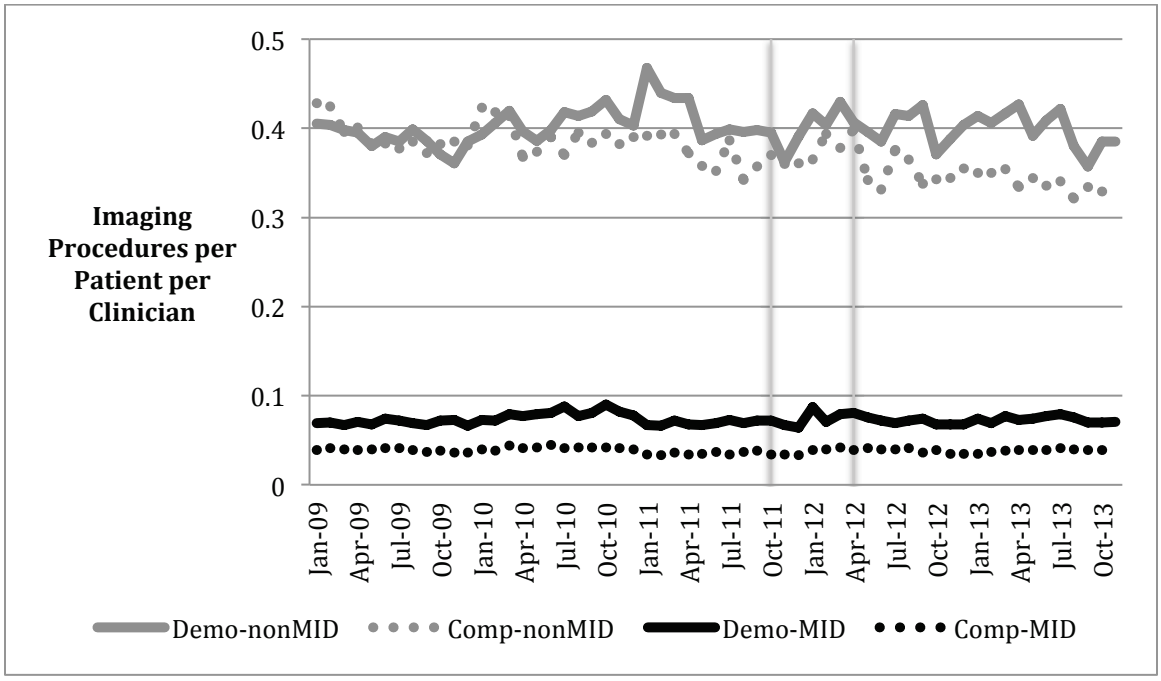
Figure 5.1 shows trends in the rate of imaging procedure utilization per unique Medicare beneficiary seen per clinician per month. The trend lines reflect the mean of the order rates in each month among ordering clinicians in the demonstration and comparison groups. The trend in MID imaging procedure utilization was roughly flat over the 2009–2013 period in both demonstration and comparison groups. The rate of non-MID imaging procedure utilization declined slightly over the 2009–2013 period.

Although trends in MID and non-MID imaging procedure utilization over time were similar in the demonstration and comparison groups, both MID and non-MID levels of imaging procedure utilization rates were higher in the demonstration group than the comparison group. This indicates that clinicians affiliated with demonstration practices were systematically different from comparison group clinicians in terms of advanced imaging utilization. We accounted for these differences in our statistical testing (discussed further in Section 5.2) in two ways. First, we propensity score weighted comparison group clinicians to more closely resemble demonstration group clinicians using predemonstration ordering volume, among other variables. Following this weighting, levels of MID imaging procedure utilization were more similar across groups, although levels were still somewhat higher in the demonstration group (data not shown). Second, we compared trends in imaging utilization over time, accounting for the residual consistent differences in MID imaging procedure utilization between groups.

The gray vertical lines in the figure indicate the initiation of the baseline and demonstration periods (note that these were the planned start dates, but some demonstration practices actually started these periods at later dates).[10] There was no obvious change in trends that coincided with the initiation of either the demonstration baseline or intervention period.

[10] These graphs show only the planned start dates of the periods. In the regression analyses presented below, practice-specific start dates were used to identify the beginning of each period.

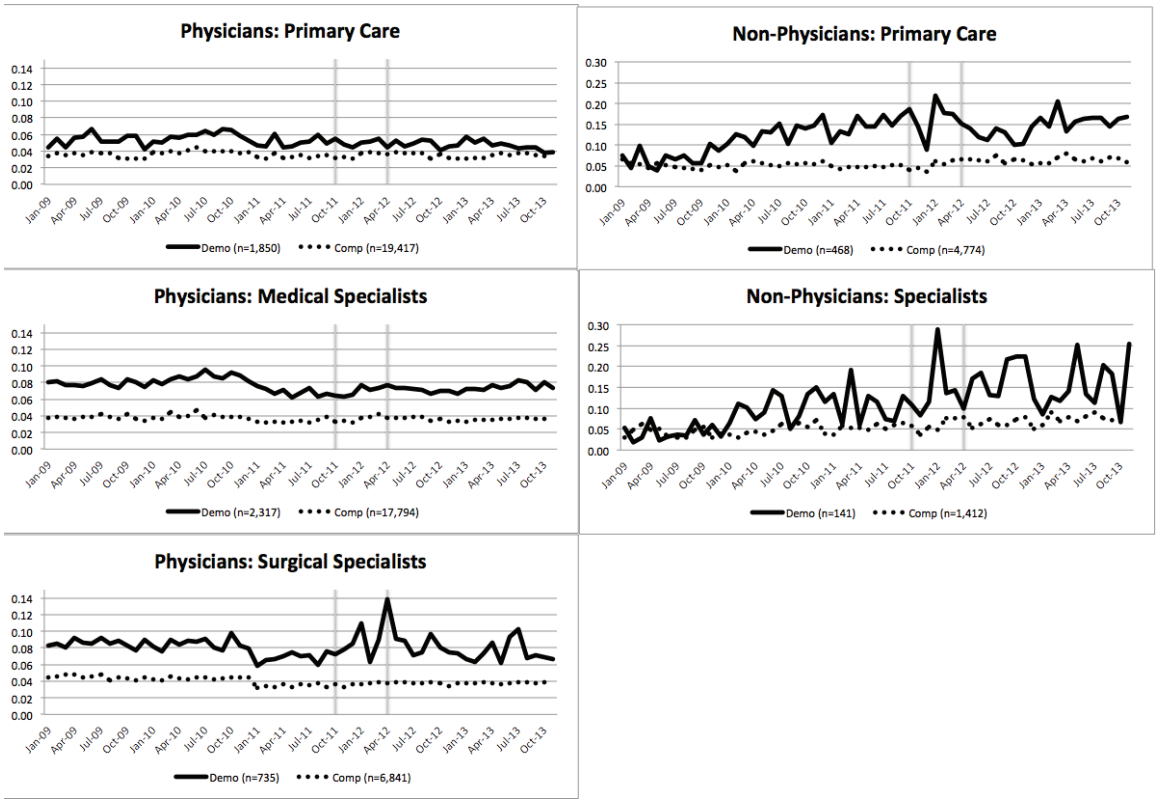
Figure 5.1. MID and Non-MID Imaging Procedure Rates per Patient per Clinician, January 2009–November 2013



NOTE: Demo = demonstration group. Comp = comparison group. The vertical gray lines indicate the beginning of the baseline and intervention periods.

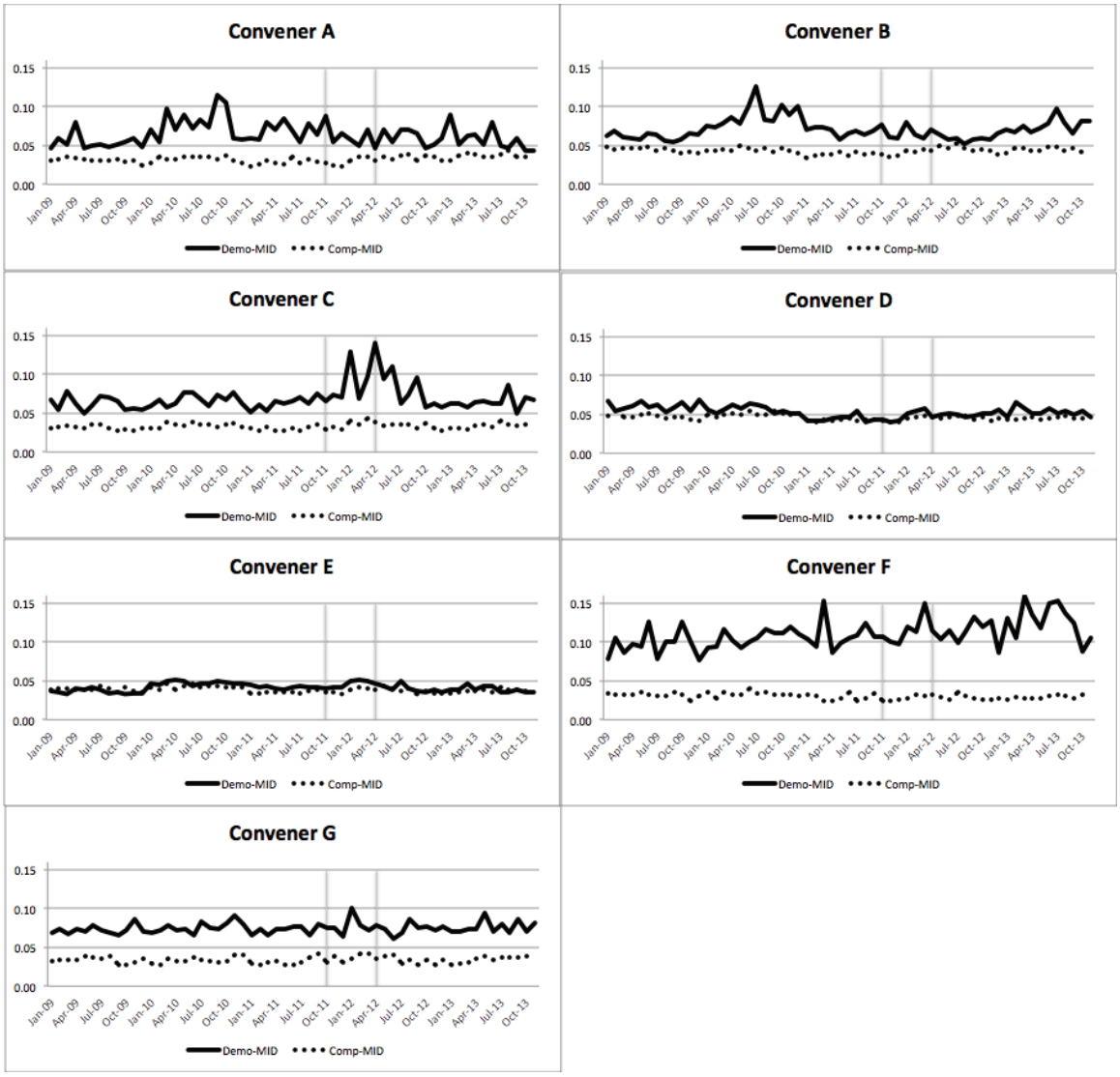
We performed stratified analyses of trends in imaging procedure utilization per Medicare beneficiary per clinician by specialty category. Among physicians, we found similar patterns of MID imaging procedure utilization across specialty categories (generalists, medical specialists, and surgical specialists; see Figure 5.2). MID imaging procedure utilization declined slightly over time in all three specialty categories, with similar trends over time in demonstration and comparison groups. The level of MID imaging procedure utilization was higher among demonstration ordering clinicians than comparison ordering clinicians in all three physician specialty categories, with the largest difference between demonstration and comparison physician medical specialists. Among nonphysicians, we found different patterns. Trends in nonphysician MID imaging procedure utilization were less stable than those observed for physicians, due in part to smaller sample sizes. Unlike physicians, nonphysicians did not exhibit decreases in MID imaging procedure utilization over time.

Figure 5.2. MID Imaging Procedure Rates per Medicare Beneficiary per Clinician, by Specialty Category, January 2009–December 2012



We also performed stratified tests by convener (Figure 5.3). Although there were some differences in trends between conveners, imaging utilization for most conveners declined over the period to 2009 to 2013. In most conveners, trends were similar for both demonstration and control group ordering clinicians, with no obvious breaks in the demonstration group trends associated with the demonstration baseline or intervention periods. The exception was Convener C, where there was a spike in demonstration group utilization at approximately the time of onset of the intervention period. The differences in levels of utilization between demonstration and comparison groups were greatest for Conveners F and G and smallest for Conveners D and E.

Figure 5.3. MID Imaging Procedure Rates per Medicare Beneficiary per Clinician, by Convener, January 2009–November 2013



5.2. Statistical Tests for a Demonstration Effect on Trends in Advanced Imaging Utilization

Although the comparison group ordering clinicians were selected from counties that were matched on a number of important variables to counties including demonstration practices (see Appendix A), we expected that there could still be substantial differences in the characteristics of the ordering clinicians in the demonstration group and comparison group. If the demonstration were evaluated by simply comparing mean values of outcome variables among the demonstration and comparison group clinicians, we would be concerned that any patterns

observed in the graphs in the previous section could be explained by lingering differences between baseline characteristics of the two groups.

We use two primary statistical tools to minimize the potential effects of confounding variables: difference-in-differences and propensity score weighting. First, we use a difference-in-differences design. By measuring trends in imaging utilization over time, this approach accounts for any consistent differences in MID imaging utilization between groups.

Although restricting focus to the differences over time for the same clinicians should account for inherent differences between clinicians, it is possible that the differences themselves depend on clinician characteristics. For example, it could be the case that a particular specialty is seeing large drops in MID image use relative to the other specialties for reasons that are unrelated to the demonstration. If there were a substantial imbalance in the relative frequency of this specialty between the demonstration and matched control sites, difference-in-differences analysis could incorrectly find a treatment effect due to confounding by specialty.

To address this concern, we performed propensity score weighting. This method (see Appendix A for detail) gives less weight in the analysis to comparison physicians with characteristics that are overrepresented in the comparison group relative to the demonstration group. Conversely, if there are physicians with a certain profile that is underrepresented in the comparison group (for example, a physician in a particular specialty with a patient population that typically has a high burden of illness measured with high mean hierarchical condition category [HCC] scores), the providers in the comparison group with such profiles will be given greater weight in the outcome analysis. In this way, propensity score weighting minimizes the potential confounding effect of observed, predemonstration characteristics.

We used regression analysis to estimate the effects of the demonstration on our primary outcome measure, MID imaging procedures per beneficiary. The measures were summed over months in each of the three periods (prebaseline, baseline, and the intervention periods), with the numerator being MID imaging procedure volume and the denominator being number of unique beneficiaries visiting the clinician in that period. In some cases where clinicians saw a low volume of patients, we observed months with at least one MID imaging procedure but no beneficiary visits due to differences between visit and imaging procedure dates. These observations have been excluded.

Table 5.1 presents estimates of imaging utilization trends between the predemonstration and demonstration periods. The results indicate that although MID imaging procedure utilization declined in both comparison and demonstration groups, there was no strong evidence that the MID demonstration resulted in a steeper decline in the demonstration group than the comparison group.

The "utilization change" column of Table 5.1 includes estimates of the change in MID imaging procedures per 100 beneficiaries between predemonstration and demonstration periods across both the demonstration and comparison groups. Between these time periods, the decrease in MID imaging procedures per 100 beneficiaries was generally small, ranging from –0.46

(Convener E) to –3.32 (Convener F). For all conveners except A, the 95-percent confidence interval does not include 0, indicating a statistically significant decrease.

The "estimated demo effect" column of Table 5.1 shows differences in the utilization trend associated with the intervention period of the demonstration. For Conveners A, B, D, E, and F, the estimate is not statistically significant (95-percent confidence interval includes 0). For both Conveners C and G, there was a statistically significant difference in trend between demonstration and comparison groups. The difference was approximately 1 MID image per 100 beneficiaries (Convener C: –1.11; Convener G: –1.29). We consider these effects to be rather small from a practical perspective. Further, both of these statistically significant results are sensitive to our decision to remove the MID images that occur in months with an observed zero beneficiary count.

Table 5.1. Estimated Change in MID Imaging Procedure Utilization and MID Demonstration Effect, Predemonstration to Intervention Period

Convener	Estimated Demo Effect (MID Imaging Procedures per 100 Beneficiaries)[a]	(95% CI)	Utilization Change (MID Imaging Procedures per 100 Beneficiaries)[b]	(95% CI)
A	0.49	(-1.21, 2.19)	-0.81	(-2.11,0.49)
B	-0.89	(-1.89, 0.11)	-0.82	(-1.42,-0.22)
C	-1.11	(-1.80, -0.42)	-1.95	(-2.26,-1.64)
D	-0.03	(-0.56, 0.50)	-0.65	(-0.89,-0.41)
E	0.48	(-0.24, 1.21)	-0.46	(-0.73-0.19)
F	-1.10	(-2.70, 0.50)	-3.32	(-3.99,-2.64)
G	-1.29	(-2.27, -0.31)	-2.64	(-3.23,-2.06)

NOTE: This analysis reports the propensity weighted difference-in-differences regression results comparing the predemonstration period that occurred prior to the initiation of the MID demonstration to the 18-month demonstration intervention period.

[a] The "estimated demo effect" column shows the treatment effect of the demonstration.

[b] The "utilization change" column shows the overall change with time from the predemonstration to the 18-month intervention period.

Next we made similar comparisons between the demonstration's baseline (rather than predemonstration) and intervention period (Table 5.2). There were statistically significant overall utilization changes between these periods for Conveners C, F, and G. Convener G had the single statistically significant difference in trend between demonstration and comparison groups that would indicate an effect of the demonstration. The mainly insignificant treatment effects strengthen the conclusion that the impact of the MID demonstration on imaging utilization was small or nonexistent. We note that the statistical insignificance of the utilization change estimates in this comparison does not necessarily mean that the overall trend of lower MID utilization has stopped. Rather, if it is a gradual process, the shorter period of time from the demonstration's baseline to the demonstration's intervention period (relative to the predemonstration to demonstration) may not be sufficient to capture meaningful differences for most conveners.

Table 5.2. Estimated Change in MID Imaging Procedure Utilization and MID Demonstration Effect, Baseline to Intervention Period

Convener	Estimated Baseline Demo Effect (MID Imaging Procedures per 100 Beneficiaries)[a]	(95% CI)	Utilization Change (MID Imaging Procedures per 100 Beneficiaries)[b]	(95% CI)
A	-0.75	(-2.04,0.54)	0.62	(-0.20,1.45)
B	-0.12	(-0.71,0.48)	-0.14	(-0.51,0.23)
C	-2.75	(-6.28,0.77)	-2.02	(-2.75,-1.29)
D	0.13	(-0.18,0.44)	-0.09	(-0.25,0.07)
E	-0.23	(-1.12,0.66)	-0.11	(-0.28,0.07)
F	-0.66	(-2.68,1.36)	-2.93	(-4.61,-1.24)
G	-1.18	(-2.17,-0.18)	-2.27	(-2.80,-1.73)

NOTE: This analysis reports the propensity weighted difference-in-differences regression results comparing the baseline period (defined as the first six months of the demonstration) to the intervention period (defined as the latter 18 months of the 24-month demonstration period).

[a] The "Baseline demo effect" column shows the treatment effect of the intervention period compared with that of the baseline period.

[b] The "Utilization change" column shows the overall change with time from the baseline to the intervention periods.

Finally, we performed a sensitivity analysis that restricts the analysis to demonstration group ordering clinicians who were identified in both DSS and claims data.[11] The results are essentially identical to those presented in Tables 5.1 and 5.2: Conveners C and G are the only to see a significant demonstration effect on utilization relative to the comparison group from the predemonstration period to the intervention period; the predemonstration period to the demonstration period changes in utilization are, except for Convener A, all significant and negative (corresponding to falling utilization overall); only Convener G has a significant demonstration effect between the baseline and intervention periods. Therefore, we conclude that our findings are not highly sensitive to the definition of the demonstration group.

5.3. Conclusion

In summary, we see evidence for a drop in utilization that is significant from the predemonstration to the demonstration period for all but one of the conveners. However, this drop does not differ significantly for demonstration versus comparison sites for five of seven conveners. For the two conveners with statistically significant reductions in the number of advanced images utilized by the demonstration compared with the matched control providers, the estimated difference is relatively small, implying an additional reduction of one MID image per 100 Medicare beneficiaries within demonstration versus control sites. Given that even these two significant results are sensitive to whether images are included in months when no beneficiaries are found for a particular provider, we consider the evidence for these demonstration effects to be weaker than the confidence intervals imply on their own. Overall, our claims analysis provides no evidence that appropriateness feedback leads to anything beyond a small reduction—if any reduction at all—in MID image volume.

[11] The control group is the same in both the main analysis and the sensitivity analysis. The main analysis includes some demonstration group clinicians who were identified in claims but not DSS records.

Section IV: Physician and Patient Experience with Appropriateness Criteria for Advanced Imaging

Section IV addresses physician and patient experiences with exposure to advanced imaging appropriateness criteria. Statute questions include:

- How satisfied were physicians in the demonstration with being exposed to advanced imaging appropriateness criteria? (Chapter 6)
- How satisfied were Medicare patients in the demonstration with receiving an advanced imaging procedure after a physician was exposed to appropriateness criteria? (Chapter 7)

This page is intentionally blank.

6. Physician Satisfaction with Exposure to Advanced Imaging Appropriateness Criteria in the Demonstration

This chapter discusses the question:

- How satisfied were physicians in the demonstration with being exposed to advanced imaging appropriateness criteria?

Clinician focus group participants reported that they were variably satisfied with the demonstration project, with some describing their support for the "idea" behind the demonstration. (See Appendix B for methods associated with clinician and staff focus groups.) However, most were not convinced that the DSS in its current form would have a significant positive impact on patient experiences or quality of care. For example, Convener A generalists and specialists noted that they liked the idea behind the demonstration, but were not sure whether ROE and the DSS in their current forms were adequately developed or integrated to be useful for clinicians. As a Convener A generalist explained:

> I would say "conceptually" I agree with [DSS]. I hope that some of the annoying features of it are worked out to a certain extent. Some of the stuff just takes more time than we want [it] to. I'm hopeful that it will be an important point-of-care resource moving forward.

Many clinician focus group participants did not believe that the available guidelines and the software linking the user interface and guidelines were well-developed enough at this time to adequately engage clinicians in a meaningful way. Thus, while they endorsed the concept of a DSS, they believed that more pilot testing should have been done prior to the implementation of a national demonstration. Below, we discuss several issues related to clinician satisfaction.

6.1. Most Clinicians Did Not Find the Guidelines Useful

Using results of surveys fielded with focus group clinicians (see Appendix B), Table 6.1 shows the survey respondents' ratings of their experiences with guidelines used with MID DSS. Overall, generalists, medical specialists, and surgical specialists were more likely to disagree than to agree with statements describing the guidelines as helpful to them. This was so for all five statements (rows A–E in Table 6.1) pertinent to clinician's perceived usefulness of the guidelines. In particular, note that row C in Table 6.1 shows that 23 percent of generalists, 16 percent of medical specialists, and 0 percent of surgical specialists reported learning "a lot" from the guidelines used to generate appropriateness ratings. In contrast, rows F and G show clinicians were more likely to agree than to disagree with statements noting that their clinician and radiologist colleagues approach advanced imaging in a manner similar to DSS guidelines.

Table 6.1 Distribution of Survey Respondent Ratings About Guidelines Used with MID Decision Support Systems, by Specialty Type

	% distribution for generalists (n=27)		% distribution for medical specialists (n=22)		% distribution for surgical specialists (n=12)	
Statement	**Agree**	**Disagree**	**Agree**	**Disagree**	**Agree**	**Disagree**
A. The DSS clinical guidelines are useful to my practice	48	52	42	58	22	78
B. The DSS empowers me when talking with patients about ordering advanced imaging	30	70	44	56	22	78
C. I learn a lot from the guidelines used to generate appropriateness ratings	23	77	16	84	0	100
D. I often read the guidelines associated with the DSS ratings	42	58	16	84	22	78
E. DSS guidelines help me to stay current with new data about indications for and against use of advanced imaging	35	65	26	74	11	89
F. Clinicians I work with typically approach advanced imaging in a manner similar to the DSS guidelines	63	37	58	42	56	44
G. Radiologists I work with approach advanced imaging in a manner similar to the DSS guidelines	54	46	63	37	67	33

6.2. Clinicians Were Not Receptive to the DSS Feedback; They Wanted More

Although conceptually interested in learning about how to improve ordering patterns, in the context of clinical practice, most clinician focus group participants reported difficulty realizing the value of DSS as it was presented. Several clinicians noted that they expected that a DSS would provide more detailed feedback that would help clinicians reduce the number of inappropriately rated orders. As one Convener C generalist put it:

> [If the intent of MID was to present the ordering physician with appropriateness ratings to influence their image ordering patterns,] I did not see that at all. . . . It didn't tell me—and if it did, I'm sorry, I never saw it—"that this is not the appropriate study for the following reasons. Here are the guidelines." So I wasn't particularly helped by this support tool as it was presented to me as a user.

Some Convener D generalists explained that their expectation was that guidelines would help them make the decision about an order as they were entering relevant information, but did not find the guideline helpful in that way. A Convener D generalist noted:

> I don't remember finding a hot key [or link] to click on a guideline that would give me an up-to-date description of a critical problem and imaging sequencing, and that's what I would expect to see. Something like a back pain algorithm

> would be "do nothing for two weeks and physical therapy for two weeks, and if you're still having a problem, maybe more physical therapy or an MRI if you had particular symptoms." So we all know those algorithms of care, but I'm not seeing that on the screen at all and maybe I just never learned where to click.

Some clinicians reported that the DSS design was not useful because it provided feedback after the order was placed, rather than before. The MID DSS is designed to allow a clinician to order a test and then characterize the reason for the order. Only after the clinician enters the order for the desired test does he or she learn whether the DSS software assesses the order as appropriate or not. Some clinicians reported that they would have preferred to enter the reasons for an order first and then receive guidance on what images to order. Explains a Convener D generalist:

> I think [the use of DSS in MID] is an approval or disapproval exercise. To me, a guideline would be something like, I type in keywords, like "right-sided back pain in a thirty-year-old female with a family history of kidney stones and a missed period," something like that. I want to know from a guideline [how] to figure out what's the best test, right? There are a number of decision thoughts there: You're going to irradiate a young female who missed a period, [or] you're going to get an ultrasound. What's the best test? I don't get that kind of decision support out of this tool. . . . I'm not seeing where there's anything to point me and say: "Order an ultrasound instead of . . . Did you order just plain X-rays first before jumping to an MRI of the LS-Spine?" I haven't seen that and so I don't feel like there's a lot of decision support.

Several participants noted that their expectations for the guidelines were not met. Some Convener B and E generalists reported that potential value associated with using an evidence-based decision support system was blunted with guideline access being available only after the order was placed. Clinicians reported that by the time the order was placed, they had already made up their mind about what to order. After the order was placed, in the context of a busy practice, they did not have time or desire to go back and revise their order. This was especially so as they were aware that equivocal or inappropriate orders could be placed using the MID system and that their reimbursement would not be affected during the demonstration.

A Convener A generalist expressed an interest in learning more from the system:

> It doesn't really do the teaching that I had hoped it would [as I was] entering into this. [I thought] you would take home some lessons every time that you ordered a test inappropriately or endeavored to do so.

Another Convener A generalist noted that the DSS did not provide the expected level of support:

> It [should have been] more like a physician's support tool, where you went down the algorithms responding to questions they ask you along the way, and [it] continually gave you feedback about how appropriate or inappropriate it was and, if not appropriate, what would be the more useful test.

6.3. Clinicians Were Concerned About the Comprehensiveness, Clarity, and Validity of the Guidelines

Some clinical areas have limited available guidelines for advanced imaging. A Convener C specialist suggested:

> In my area of practice, there aren't real clear guidelines for ordering tests. That's why, when I said previously, that I don't get much kickback because of the subpopulation I work with, there just aren't nice guidelines for what's appropriate and what's not.

Many specialties simply lack guidelines for the clinical scenarios they routinely encounter. A Convener D specialist explained:

> From a cardiology perspective, our group tries to be consistent with following the guidelines, but there are certain areas where we don't have guidelines. [In these instances,] our clinical assessment is more important than the guideline, . . . so I think the clinical situation is the most important thing. When we are not sure, then probably guidelines will direct our course, but that is how I think we try to use the guidelines. Sometimes we don't have guidelines because the evidence just is not there.

A related challenge concerns the DSS design, which focuses on the procedure to be performed (e.g., a scan of the head) rather than on a particular medical condition that might be associated with a guideline. According to a Convener C generalist:

> One of the biggest problems was that, at the end, you get this message that [your order] is not covered by guidelines. I think part of the reason was that all the decision support rules that seem to have been made were based on the procedure itself, rather than a medical condition. The logic was based on, okay, a CT of the head, what indications, what output to give, whereas it should probably be more based on a condition. So if I'm trying to evaluate lower back pain, then what the logic would be? So it was kind of a backwards way. The guidelines were based on procedures, which is also one of the reasons why, when we went through the initial testing and even the folks were trying to validate the system in the beginning, they were giving very, very specific indications then, and naturally it would work for those indications, but that's not the way clinicians actually practice. You already made up your mind that you're trying to evaluate a certain set of differentials and you would like the system to, first of all, capture the indications, and then apply guidelines based on the indications, rather than on the test that you selected.

Some specialists stated that there are numerous discrepancies between radiology guidelines and specialty guidelines, as well as with local institutional practices. These include discrepancies between guidelines from the ACR and other specialty societies in some areas. These differences require the DSS system to choose one option or the other:

> I haven't looked at the following situation, but I can imagine where there's conflict, and that is using CT scanning for screening for lung cancer. So the

> guidelines are currently being developed, and there's a fundamental philosophical conflict between the NCCN [National Comprehensive Cancer Network] and the professional societies, and which patient populations to screen. So if the DSS system is going to provide guidance, a decision has to be made whether to provide both sets of guidelines or show a bias towards one.

Explained a Convener C specialist:

> The inputs to say what is appropriate or not appropriate actually do not correspond to published cardiology guidelines . . . For example, . . . important variables that are decision trees in the cardiology guidelines, like whether or not the patient can exercise, are actually not in the form.

Similarly, a Convener C generalist noted that radiologists observed "a lot of discrepancy between what the guidelines contained and what the institutional standard for doing things is."

6.4. Many Clinicians Reported Not Seeing and Not Viewing the Guidelines

MID DSS was intended to pair concurrent feedback of the appropriateness rating with a real time educational intervention. This was to include a recommendation for an alternative image order with a more appropriate rating for the stated clinical indication if one were embedded in the guidelines. Additionally, real time access to the guidelines was expected to be a component of the DSS intervention.

A Convener E generalist commented on the importance of these educational interventions:

> The greatest potential would be in terms of educational feedback and trying to improve the ordering process if the test actually wasn't approved or didn't go through until it either met the appropriateness criteria or there was education to change the test, if that were appropriate.

However, in practice, DSS guidelines were not readily accessible to ordering clinicians.

Although most participating clinicians were aware of DSS appropriateness ratings assigned to their orders, the majority reported that they had not seen the guidelines used in DSS. Explained a Convener C generalist:

> In my practice, I cannot remember a single time when the guideline came up and actually gave me information about a different test that I might consider or anything that gave me information that might be useful. What I recall is that I either checked the right boxes to let it order the test that I thought was necessary, or I specifically recall the very bothersome times when I had a critically ill patient and needed a scan immediately, but sat there for minutes and minutes trying to put in truthful answers to get the thing to accept a scan, but it never got to the point of even saying, "This isn't reasonable because of X, Y and Z," or "Consider this;" it just caught you in a quagmire of questions where you could not get past them.

Many clinicians who received ratings of "equivocal" or "inappropriate" stated that they either had not noticed or did not remember seeing links to guidelines or, if they saw a link,

did not click on it to review the guidelines. Both generalist and specialist clinicians noted the format of the guidelines was not compatible with reading within the context of a time-limited clinical visit with a patient.

A Convener A generalist explained:

> I was already uncertain [whether the image should be ordered with or without contrast] at the time I ordered the initial test, and when it [DSS concurrent feedback] recommended the order be placed without contrast for that particular study, I didn't go to the guideline to verify the rest of the information. Although I noticed where the guideline was, I did not actually look at it."

For this clinician, receiving the appropriateness feedback was adequate to influence the selection of the order without contrast even without review of the guideline.

A common concern from clinicians was that the guidelines, when accessed, are too long and hard to follow. Repeatedly, both generalist and specialist clinicians described the fast and often intense pace of their clinic sessions. Time was often a major factor, as they wanted to avoid patient waits and also wanted to be prompt for the scheduled professional activity that followed their clinic. Additionally, clinicians often noted the need to stay on time within appointments since urgent clinical problems often arose but were not allocated adequate time. Thus, even a brief delay associated with ordering an image or accessing guidelines could frustrate a clinician. Clinicians often remarked that in a busy practice, even a three-minute delay for each patient can cascade into an extensive delay for patients scheduled later during the clinic session.

A Convener E generalist documented the challenge associated with reading DSS associated guidelines during the clinical encounter:

> I pulled the ACR appropriateness criteria on cerebral vascular disease . . . and the first 11 pages are different scenarios or variants that you have to kind of sort your way through to even find out what you're looking for, which is quite a lot of pages to sort through and read through to even pick the scenario. And then it's followed by another ten pages of literature summaries, and it talks about backgrounds and imaging, and thrombolytic therapy. It's a very elaborate and probably informative didactic discussion, which, if you have an hour or two, you may learn some things from. But in terms of it actually giving you a very readily usable and easy guide to decisionmaking that's going to be fitting into the workflow of a work day, it's not even close to that.

A similar perspective was provided by a Convener B generalist:

> It's an eight-page PDF from the ACR that basically goes through—the first two pages are a chart and then after that is a three-page PDF. . . . Basically, it's a summary of literature to help support why you should get it or not, and so that was the eight pages you do have access to. . . . I would agree with my colleagues, it's not very practical, either put there or set up for real-time use.

Only two of 97 focus group participants stated that guidelines were useful or helpful. One participant liked the guidelines and recommended that all clinicians familiarize themselves

with radiology guidelines. Another noted that while s/he likes the guidelines, s/he was not convinced that the radiologists follow them.

6.5. Actions Taken When Orders Are Rated Equivocal or Inappropriate

When an order is rated equivocal or inappropriate, clinicians may change, cancel, or retain the order as is without or with changing the reason for the order.

6.5.A. Changing or Cancelling the Advanced Image Order

Section 2.5 discusses rates of cancellation and change noted during the intervention period for images initially assigned an inappropriate rating. No convener changed modalities (CT to MRI, MRI to CT, or advanced imaging to a non-MID procedure) or changed contrast status more than 11 percent of the time when the initial order was rated inappropriate. Conveners were more likely to change their initial order when receiving an equivocal rating, but only marginally so.

Table 2.6 shows when an order is deemed equivocal or inappropriate, clinicians may elect not to order an originally selected test and order an alternative radiologic or nonradiologic alternative instead. An important feature of MID's intervention period was the pairing of the feedback of an inappropriate or equivocal rating of an order with the display of an alternative imaging procedure, should one be recognized within the guidelines as having a higher rating than was assigned to the image ordered. Table 2.7 showed that, across MID images, just under 40 percent and nearly 80 percent of inappropriate and equivocal orders, respectively, were associated with feedback of at least one alternative image type.

Substantial variation was noted across image types. While alternatives were available for at least 25 percent of inappropriate and at least half of the equivocal images for all procedures except SPECT-MPI, alternative images for these cardiac procedures were provided in association with feedback reports for fewer than 3 percent of images. Explained a Convener C specialist:

> In our center, [we cannot] order the test except through the EHR [electronic health record], so we just run into a wall. [If our order is inappropriate], we just don't order the test; we try and order an alternative test, even though that is not the test we would have wanted or feel that is appropriate. So, for example, for a [cardiac] nuclear stress test, if we just hit that wall, then we order an exercise treadmill test because we cannot place the order. We don't have a back door to place the order.

6.5.B. Changing or Cancelling the Advanced Image Order

Section 2.5 also documents that clinicians rarely canceled orders that were rated inappropriate during the intervention period. The only exception was Convener D, whose clinicians canceled up to 18 percent of inappropriate and 8 percent of equivocal initially rated orders. As documented in Section 2.5, clinicians were more likely to change an order if an

alternative procedure with a higher appropriateness rating than was assigned to the initial order was recommended.

6.5.C. Retaining the Advanced Image Order

Clinicians may go back, try to find what causes an inappropriate rating and change either symptoms or diagnoses hoping that the order would be rerated as appropriate. Explained a Convener D generalist:

> It's happened several times for sure, where my initial orders have come back [inappropriate], and then I've gone back and tried to figure out why, so rather than submit, I'll try to look at the clinical scenario, or signs and symptoms, or ICD-9 code and figure out if it's the right one. Sometimes if I figure out, well, if I select this one and redo it, it comes back appropriate. It's not that I'm gaming, it's just that I'm trying to find out a more appropriate ICD-9 code.

Clinicians may ignore inappropriately rated orders; they override such orders by signing the attestation that they understand they are ordering a study that has been rated as inappropriate. All DSSs in the demonstration were programmed such that ordering clinicians must attest that the order they place is as they believe it should be, regardless of the assigned appropriateness rating. As explained by a Convener C generalist:

> I hate to say it, but I think that people learned what box could be checked to get the test when you wanted it, and people just checked that box whether it was a truthful answer or not.

Other focus group participants reported overriding inappropriate orders automatically to expedite the process of advanced image ordering. As a Convener D generalist explained,

> I've got to admit—usually, when I'm doing this, I'm in such a rush . . . I'm just trying to get it done as quickly as I possibly can and keep moving. While it's a tool that in theory would be great, the reality is, when I'm in the room, I'm not thinking about learning, I'm just thinking about getting through. So when I end up doing this tool, I end up finding that I click on what I refer to as "the cop-out" most of the time, and then I just end up clicking the box that says you've clicked on "the other," so just click over here to say you accept it.

Explained a Convener B generalist, "We just use our best clinical judgment, take every individual patient, and then go on."

As shown in Figure 2.5, across all conveners, clinicians changed only between 4 and 8 percent of inappropriate orders, meaning that for most orders that were initially rated as inappropriate during the intervention period, clinicians continued with their initial order. Figure 2.6 shows the prevalence of changes among initial orders rated inappropriate was fairly consistent across image types, ranging from approximately 3 to 8 percent.

The most frequent reason clinicians provided for retaining the order as originally placed even when the rating was equivocal or inappropriate related to clinical judgment about the patient for

whom the order was written. Convener C specialist provided an example of this most frequent reason for retaining an order, even when it was assigned an inappropriate or equivocal reason.

> I did have one case where I was imaging for a cerebral aneurysm. And I had sort of a low suspicion in an elderly patient with heart disease, and they recommended computed tomography angiography (CTA), which, of course, is the best study. But I didn't want to put the patient through that, in the sense of cardiac overload and complications. This was an [older] person, and I had low suspicion. So I just overwrote it, [continuing with the originally ordered image despite an inappropriate rating].

The clinician ordered the desired test and signed the attestation for this elderly patient believing that, based upon the patient's frail status and comorbidity, the best test was other than the one recommended in the guidelines.

Other clinicians noted a commitment to their own specialty guidelines as more important than following radiology guidelines. A Convener D specialist explained,

> I think most specialists are going to practice within the guidelines within their specialty anyway, and I don't think many of us are going to have a problem ordering and using the system, but there's an occasion, I think, where you might have a concern about what study to order—and if I think I know what I want, I'll go ahead and override it and then attest to it, but other times I call the radiologist and they walk me through it.

Other clinicians noted that the best strategy for progressing efficiently through their busy clinical schedule was to respond to an inappropriate or equivocal appropriateness rating by continuing with the order without any change. This reflected the time saved if clinicians could avoid changing orders.

6.6. Clinicians' Concerns About Time Required to Place and to Change an Order with DSS

Table 6.2 shows that across conveners, the mean time clinicians estimated to place an advanced image order was longer at the time of the focus group than it was reported to have been one year prior to MID implementation. The reported times increased initially after Phase I MID implementation as clinicians became oriented with the new protocols for advanced image orders. Reported ordering times were shorter during Phase II than during Phase I, presumably because clinicians had become accustomed to ROE and the DSS. However, clinicians described orders involving changes in response to DSS feedback as the most time-consuming component of MID. When asked at the time of the focus group about clinician time involved in completing average orders compared with time spent prior to MID, the mean across conveners increased by 3.3 minutes with the median (not shown) increasing by three minutes.

The extra time required to enter image orders seemed to be particularly frustrating for clinicians who care for patients with multiple comorbidities, as these clinicians have to spend

more time with their patients. In the words of a Convener C geriatrician, "I'm in a geriatrics practice, so things take a little longer. And that extra time really just added to slowing down my workflow."

Table 6.2. Clinician Reported Time (in Minutes) to Complete MID Image Order, By Convener

	1 Year Prior to MID Implementation	During Baseline MID Period	During Intervention MID Period, When Using Both ROE and DSS		At the Time of the Focus Group
Clinician Focus Group Sample Size	Mean (SD)	When using ROE, but prior to initiating DSS Mean (SD)	If not changing order in response to DSS feedback Mean (SD)	If changing order in response to DSS feedback Mean (SD)	On Average Mean (SD)
9	2.1 (1.3)	2.8 (1.9)	3.6 (2.1)	4.0 (1.5)	5.1 (2.2)
11	2.4 (1.6)	2.3 (1.7)	1.9 (1.3)	3.4 (1.4)	3.7 (1.2)
14	3.0 (2.1)	3.0 (1.8)	2.9 (1.4)	6.6 (3.4)	8.1 (5.3)
16	5.9 (8.2)	5.6 (7.0)	4.8 (4.8)	7.6 (5.1)	9.2 (13.8)
8	5.0 (3.7)	5.9 (4.8)	4.1 (3.6)	7.3 (6.6)	8.8 (8.9)
All (n=58)	**3.9 (5.1)**	**4.1 (4.7)**	**3.6 (3.5)**	**6.0 (4.5)**	**7.2 (9.0)**

NOTE: Across conveners, no significant difference was noted comparing time to complete an advanced image order from one year prior to MID implementation to during the baseline MID period when ROE was used but no DSS feedback about appropriateness was received by ordering providers. Similarly, no significant difference was noted between the baseline period and the intervention period if clinicians were not changing an order in response to DSS feedback. However, as shown in the *All* row, the mean time to complete an order during the baseline period is 4.1 minutes (SD=4.7) compared to a mean time of 6.0 (SD=4.5) minutes during the MID intervention when the order is changed in response to DSS feedback (p=0.0063). Similarly, there is a substantial increase in the mean time to complete an order on average at the time of the focus group compared with during the baseline MID period (mean time difference 3.1; p=0.0037).

6.7. Clinician Perceptions of the Impact of Guidelines on Ordering and Clinicians' Relationships with Others

6.7.A. Impact on Advanced Imaging Order Appropriateness

Clinicians varied with respect to their views about the impact of guidelines on ordering. Most clinicians reported the belief that the demonstration had not had a significant impact on their image order appropriateness or volume. As an explanation for the lack of impact, some noted that they were already conservative in their ordering practice. As a Convener D specialist stated,

> I don't think it [DSS] has really impacted [us] because for a long time, we have already been kind of conservative in ordering the images. If you're asking whether it has resulted in less ordering or more appropriate ordering, I don't think it has changed a whole lot.

Despite this, some clinicians reported their belief that the appropriateness feedback that clinicians receive makes them think more carefully about each order they place. A Convener A generalist explained:

> If anything, [DSS] may have made it [our advanced image ordering] slightly more towards appropriate, just because I'm thinking about it more. Maybe that's just me being hopeful, but it may have swayed it a little bit that way.

6.7.B. Impact on Clinicians' Relationships with Specialists

Traditionally, one reason that generalist and specialist clinicians refer patients to a specialist or subspecialist is for advice about the value, timing, and specification of advanced image orders. Some focus group participants documented an increasing propensity to refer to a specialist following MID implementation. Reasons cited for their changed behavior included the time burden associated with using the DSS, frustrations with the guidelines (especially inconsistencies between MID guidelines and local guidelines), and displeasure at having their clinical expertise evaluated with explicit quantitative ratings that could be recorded for others to see.

One Convener D generalist explained a preference to defer to a specialist on advanced image ordering:

> If you know you're going to send somebody to neurology anyway, you'll maybe not order the test because they will. Then, they'll get bagged with the utilization rather than you.

Some generalists felt that specialists were more aggressive in ordering images. In the words of a Convener A generalist:

> I know if a patient has a headache or a neurological complaint, if I send them to the neurologist, they'll probably end up with an MRI of their brain and their whole spinal column. I may not be as aggressive about it, but I can pretty much—OK, we're going to pan the image. I don't know if that's appropriate or what, but that's been my experience. The specialists tend to be much more aggressive with the advanced imaging.

Generalist clinicians sometimes reported referring patients to specialists instead of risking placing a potentially inappropriate order and "getting dinged." However, some Convener A specialists suggested they are now ordering *fewer* tests because they see a lot of patients who already come to their offices with multiple imaging tests already having been completed.

Not all clinicians experienced increased contact with specialists. For example, Convener C clinicians had a different perspective on this issue.

> I think we are so used to figuring out how to get it done on our own, that if we think we want to CAT scan, we do what we need to do to get through to get that done. I don't think that most of us are changing and just going: "Well, I'll refer them to GI instead."

6.7.C. Impact on Clinicians' Relationships with Radiologists

According to focus group participants, clinicians generally did not think that this demonstration project had any significant impact on their relationships with radiologists. Nonetheless, clinicians suggested that ROE and the DSS involve electronic transmission of data. This method differs substantially from the traditional method of information transfer that involves patients hand-carrying the order on a piece of paper from their clinician to the radiology department. The rate of loss of information is substantially less with the newer methods. As a corollary, clinicians believed that their offices and the offices of radiologists may have to invest less time in placing clarification calls about scheduling and clinically relevant information about the reason for the order. Several clinicians suggested the reduced number of required calls increases the efficiency and camaraderie of interactions between clinicians and radiologists.

6.7.D. Impact on Clinicians' Relationships with Staff

In many practice settings, staff play an important role in the advanced image ordering process. Focus group clinician and staff participants agreed that staff seemed more involved with advanced image ordering for specialist than for generalist clinicians. They attributed this pattern to the smaller number of image types that specialists order compared with generalists (see Figure 3.2). Both generalist and specialist clinicians believed staff would be more likely to serve as a reliable and accountable help in ordering advanced images if they could understand the nature of the test and the reasons for the orders. They would most likely occur in a setting where clinicians focused on a few clinical conditions involving the ordering of only a few image types. Staff are more likely to master the ordering process when they are responsible for ordering a smaller number of image types for a more well defined set of reasons. While several specialists reported satisfaction delegating to staff some aspects of image ordering with the DSS, such satisfaction was expressed far less often by generalist clinicians.

Ordering staff and clinicians reported similar experiences with ROE and the DSS. Staff also felt that although ROE may be useful for tracking orders, the DSS was time-consuming with little benefit, and had a negative impact on staff workload.

7. Medicare Patient Satisfaction in the Demonstration with Receiving an Advanced Imaging Procedure after Physicians Were Exposed to Appropriateness Criteria

This chapter will discuss the question:

- How satisfied were Medicare patients in the demonstration with receiving an advanced imaging procedure after a physician was exposed to appropriateness criteria?

In Chapter 6, we provided results on physician satisfaction with the demonstration, drawing upon results from clinician focus groups held with several conveners. In this chapter, we shift the focus to patient satisfaction, reporting on results from both the physician focus groups and from two patient focus groups. Participants in the patient groups were adult Medicare beneficiaries who had one of the 12 MID advanced images during the previous three to six months. A total of 13 patients participated.

7.1. Physicians Felt the DSS Implementation Did Not Have Any Substantial Impact on Their Relationships with Patients

According to physician focus group participants, the DSS implementation did not have any substantial impact on clinicians' relationships with patients, patient satisfaction, or quality of patient care. Most clinician focus group participants agreed that their patients were not aware that physicians were using a new protocol when they used ordered advanced images.

Table 7.1 shows clinician focus group survey respondent reports of the impact of MID DSS on patient experiences. (See Appendix B.) Overwhelmingly, clinicians indicated that Medicare patients are not aware that their clinicians are using the DSS to order advanced imaging procedures. They also indicated their perception that most Medicare patients are satisfied with receiving an advanced imaging procedure without or with a clinician using a DSS.

Table 7.1 Focus Group Survey Responses About the Impact of MID DSS on Patient Experiences

	Do you believe Medicare patients are aware that their physicians are using DSS to order advanced imaging procedures?			How satisfied would you say Medicare patients are with receiving an advanced imaging procedure after a physician uses a DSS?	
Respondent	**% Yes**	**% No**	**% Don't Know**	**% Unsatisfied**	**% Satisfied**
Total (N=61)	7	83	10	20	80
Generalist (n=27)	7	81	11	19	81
Medical Specialist (n=22)	5	81	14	17	83
Surgical Specialist (n=12)	10	90	0	29	71

As shown in Table 7.1, most clinicians believed their Medicare patients were not aware that they were ordering advanced imaging using a DSS system. In the context of most focus group participant clinicians reporting their patients were unaware of the DSS, these clinicians noted that most patients were satisfied receiving an advanced imaging procedure.

7.2 Physicians Also Noted Two Potential Areas of Impact in Their Relationships with Patients

Although clinicians did not report any significant changes in their relationships with patients, two potential impacts of the DSS on patient-provider relationships were suggested. The first concerns a potential disconnect during the clinical encounter between clinicians and their patients as clinicians using a DSS spend time facing the computer to order images. The second pertains to the potential for a positive benefit on the relationship by helping clinicians to explain why an order a patient requests is inappropriate and thus not needed.

Clinicians who order images while patients are in the room worried that electronic image ordering might send the wrong impression to patients, who might think their doctors pay more attention to managing their charts than to addressing the patient's concerns. As a Convener D specialist explained,

> I'm sitting with a patient and working on the computer, it is an absolute negative satisfaction for a patient. The patient is saying that these days, doctors are more on the computer than talking to the patient.
>
> It takes me a long time, and I'm clicking in the computer all the tests. I'm putting in the medication and consultation links, and all that. I spend a lot of time in front of my computer while the patient is sitting there and kind of waiting for me to finish all the computer work. So I think [that while ordering tests on computer is] good because it's streamlined, it helps to be electronically savvy, but, in my opinion, at the same time, it kind of takes away from the time with the patient, one-on-one.

To address a concern about reduced face-to-face contact with patients during the image ordering process, some clinicians position their computers in such a way that helps them look at the monitor and the patient at the same time. Other clinicians explained that they try to be creative and use their sense of humor to explain to patients why it takes so long to order an image.

Another potential impact of the demonstration described by clinicians relates to the use of DSS to positively affect their relationships with patients by helping them explain the reasons that some orders requested by patients may be inappropriate according to national guidelines. Some clinicians noted that sharing evidence from the national guidelines could help the clinician explain to that patient that there is no need to perform the requested image. As one Convener A specialist stated:

> When patients come in with a headache, and, you know, the doctor is talking to the patient, and the patient is demanding an MRI scan, [it's] helpful to have the support to say: "You know, your symptoms don't support this." It gives [clinicians] a leg to stand on in that area, some legitimacy. So I think that it would help.

Multiple clinicians acknowledged this *potential* for the DSS. However, several clinicians felt that the DSS they used was not robust enough to support clinicians in steering patients away from unneeded advanced images. As a Convener A generalist explained:

> In an ideal situation, you could use [the DSS] to steer [patients] away from the [unnecessary] MRI they were demanding. And you could sort of rest upon this decision algorithm that was vetted with experts so that you could point that out to the patient. This could improve things. . . . [But] the only thing they [patients] see is me sort of fumbling through another interface and trying to get the order entered, you know, having to take a little more time doing that.

A few clinicians noted that the demonstration contributed to patient understanding of what physicians do. For example, several Convener D clinicians mentioned that having patients observe how doctors order their tests was a positive change:

> I think sometimes it could be positive—in the sense that, in the past, we used to do the same kind of work, but the patient was not seeing us. Now when the patient is sitting with me, and he's actually seeing that I'm doing these things, he knows that this is the additional time I'm spending on his visit. . . . I think to a certain extent the patient feels that's OK, you're spending some time for the patient's own care there.

Some clinicians suggested that they were uncomfortable discussing the impact of ROE and a DSS on patient's quality of care without seeing the results of this evaluation. As one Convener D generalist put it, "Do I think it's affected quality? I don't think it has, but until you measure and count, you really don't know."

7.3. Patients Were Not Aware of the Demonstration

Participants in the patient focus groups also indicated that they were not aware of the demonstration. All patients in the focus groups reported that their doctor directly ordered the advanced image test using a computer. But when asked whether their doctor consulted with other doctors or used a computerized tool to help them decide which test to order, the patient participants stated that they did not. None of the patients in the focus groups seemed to be aware of or familiar with the MID DSS.

Further, none of the patients reported noticing a reduction in "inappropriate orders." None of the patient focus group participants reported having to wait to get an advanced image authorized and none reported experiencing delays in getting an advanced image scheduled and/or performed. In fact, some participants reported that their doctors did not have to obtain authorization from Medicare or Medicaid to order an advanced image. One male participant said, "Once your immediate attending physician decides you need a CAT scan, it's my experience that ordering is immediate."

The majority of patient focus group participants reported receiving help from the doctor's office in scheduling the advanced image and having had the image done in a timely manner (usually within a few days to a week). As a male participant put it, "They call you and they ask you what day is—to see what days you have available. You don't wait."

Patients also reported that they were given a choice about where to have the advanced image done.

7.4. Patients Expressed a Generally Favorable Attitude Toward Physicians' Use of Computers

Although some physicians expressed concern about the potential to spend too much time at the computer during a patient encounter, participants in the patient focus groups were generally favorable toward their doctors' use of computers. All patients reported that their doctors use computers to pull up the patient's medical record, review the patient's history and test results, record notes during the visit, order tests, and in some cases, order prescriptions for the patient. Some patients reported that their doctor uses their computer to pull up and print an image from their test result. In addition, some patients reported that their provider uses a computer to look things up on Google and print off information they want the patient to have.

> And a lot of times they're typing stuff and they Google and they give you a paper that tells you know, the same thing that you could pull off at home but you wouldn't know what to look for. So they do use the computer.

While some patients reported that the consultation room has a computer, others reported that their doctor carries a tablet computer with them as they go from one patient to the next.

Many focus group participants felt that doctors' use of computers had benefits. Some participants reported that they liked seeing their doctors' record notes from the visit while it was still fresh in their mind. Patients appreciated that doctors had access to the patient's entire medical history through the computer, that they were able to order tests and receive the results, and, in some cases, submit prescription orders for the patient to the pharmacy of their choice—all through their computer. Several focus group participants reported that their doctors provided a printed summary of the visit for the patient at the end of the visit.

> I don't know of other places, but at my clinic over the last two years, before you leave that doctor's office, you get a summary of what transpired that day and a history of the medications you're taking and it's wonderful. It tells you what appointments are coming up, what appointments you need to make, and what blood work you need to do.

A few participants, however, expressed concern about doctors' use of computers. Some patients reported that they did not have much choice in the matter. Explained a male participant:

> But the issue of the new way of constantly, of keeping track of everything is they're trying to type, they're trying to do this, at the same time paying attention, and pick up in a very short period of time what's going on with their patient. It's a conflict, without a doubt. And now they're all running around with pads.

Another participant explained that computers can interfere with communication between the patient and the provider:

> And he'll say, "Hold on a minute, let me type this in." So you just have to sit there for a minute till he gets through it, and then ask another question. But that's kind of the sign of the times.

However, another participant reported having a very different experience of the doctor's typing into the computer:

> The typing has been great. . . . He's typing or writing and is always in connection. I'm in computers. He connects with my pharmacy. He's connected with the medicine I have. That computer keeps up with everything that he's doing and also he's reporting, because even after we're sitting there, he has our consultation and then he put a report, okay, we switch to medicine, so you want to take this medicine, break it down, how I should take it. All this stuff before I get out of the seat.

Generally, patients expressed a favorable attitude toward doctors' use of electronic health records (EHR). Several participants mentioned that their physician's practice had transitioned to using an electronic medical record. In general, focus group participants stated that they like the fact that all their health information is stored in one central location and that it is available to the different providers they see, including doctors who provided care when they visited an emergency room. Several participants stated that they want their health providers to be

able to access patients' medical records even when they are receiving care when they are out of state:

> If I'm in California and get hurt and there might be some medication that's harmful to me . . . my records already let them know. That doctor that's working on me there, hopefully he can pull that chart up in my name and boom, as opposed to giving me something that would be detrimental to me, he knows just like he's my personal doctor.

While some participants expressed concern about the fact that their health information was available through the Internet, others emphasized that their medical record was secure and required a password.

Some of the participants reported accessing their EHR through their clinic's patient portal. Focus group participants who had used it reported that they found it very convenient to have information on their medications, upcoming appointments, and test results available from their computer. Explained one participant:

> You know, and it is good, because our society has changed where we used to have to wait for results. We're like a microwave popcorn society now, that you want to see what's going on right now, as [opposed] to I'm going to wait two and three weeks and . . . your mind messes you up, because then you start wondering, "well what's going on, what's that?" Now you get it right there. We don't really have to wait.

7.5. Patients Emphasized the Importance of Advanced Imaging for Their Care and Also Expressed Value in Communications with Physicians About Imaging

Physicians had expressed the hope that the DSS might positively affect their relationships with patients by helping them explain the reason that some orders requested by patients may not be considered appropriate according to national guidelines. Patients did not specifically comment on this issue (since none were aware of the demonstration). However, many described their communications with physicians and emphasized the usefulness of advanced images and the results of these tests in providing information that was important for patients' health.

Most focus group participants reported that their doctor had explained the purpose of the advanced image before they ordered it and also explained how the test would be conducted. Explained a participant:

> There's a lot of interpretation to narrow down what I actually have going on, how advanced it might be, and so forth. So a CAT scan, in fact, being exceptional imaging, would give them an idea of the severity of the scarring and so forth in my lungs. So I knew what I was going in for and why.

Many patients expressed confidence in the data that their advanced imaging would provide. A participant explained:

> I knew why he was ordering it—they saw something in my back and they would need to find out what was really going on because I could hardly walk sometimes. When I'd get up, I couldn't get out of bed, so they said they were ordering an MRI and that would tell everything.

Patients also described the ways they felt physicians would use results from the advanced image tests. Many reported that their doctor used the test result to either diagnose the source or severity of a health problem (e.g., back pain) or in preparation for surgery. Explained a participant:

> Strictly as diagnostic, to see to what degree your condition is, whether it's progressive, or whether it's bigger or littler, or it's spread, or whatever. And that is a, "we want to do this because then we'll learn more."

Another participant stated:

> I think they can make a comparison. Let's say you and I, you had an MRI, I had an MRI, they can make comparison between those two tests, look at them and we're just doing a study, really. I'm studying your MRI, and I'm studying Scott's MRI. And let's see, what's the difference.

Yet another participant explained:

> My doctor uses the results to check the progress of my tumor, to see how much they've grown over a certain amount of time and whether they need to be taken care of or not.

In addition, participants reported that their doctor reviewed and discussed their test results during a follow-up visit with their provider.

7.6. Patients Emphasized Both Physicians' and Their Own Roles in Decisionmaking for Advanced Orders

Patients tended to emphasize the individual physician's role in deciding whether to order an advanced image. Focus group participants were asked whether they thought another doctor would order the same advanced image test their doctor had ordered. Participants stated that it would depend on the doctor. While some participants reported that they had seen several doctors and all had said that the ordered test was the only option, others stated that the doctor's recommendation on whether to order an advanced image depended on their knowledge and expertise. Explained a participant:

> [My orthopedic surgeon] avoids the cost of the CT-scan but it's a special type of ultrasound with specially educated technicians that can actually read the tear and where the problem is in the rotator cuff. And yet I've known many, many people who say, "How can she do that without a CT-scan? The CT-scan is what helps the diagnosis." She avoids the cost of the CT-scan, using a procedure that's less invasive, but you have to have skilled and trained technicians looking for the proper thing with this type of ultrasound. She's never failed.

Another participant explained that some doctors order tests to make more money but that this was not the case at the practice they visit, "where there aren't any wasted operations, you might say. Because at other hospitals we'd see that over one-half of all operations are not necessary." Another participant joked: "They're trying to make some money . . . he's got a boat payment coming up." And yet another participant explained that some doctors order less expensive tests first and then order more expensive tests only when necessary:

> I think that, depending on the doctor, and I'm not saying they're not all qualified, but sometimes they want to take the least expensive route. We start at level one, therapy first. If that doesn't work, we move up to the next position, the next thing.

Overall, patients emphasized that physicians exercise their own discretion in deciding whether to order an advanced image. Patients did not seem familiar with the idea that guidelines exist or are associated with appropriateness ratings that might influence ordering of images.

Many patients also described their own roles in the decisionmaking process. When asked whether their doctor had discussed treatment options with them and their families, focus group participants reported a range of experiences. While some participants reported that it had been their own personal decision to have the advanced image, others stated it was a joint decision with their provider, while others reported that their doctor explained that the test being ordered (e.g., MRI), was the only option for their condition. Explained a participant:

> It was [for] myself and [it was] my own decision: my life, my decision. Yes, I want to know and . . . you can't get information on the surface if it's an internal [problem]. I would never deny myself something that may be informative . . . because what happens to me is, my son is going to be affected as well. It's our history, family traits.

Another participant explained that decisions depended on the situation: "Well, I think it's according to your situation. Because with my back situation, that's the only thing that the doctor can really find what's going on is [with] an MRI."

Only one participant reported that their doctor had not involved them in making the decision about having an MRI and she subsequently switched doctors.

> I don't think I needed an MRI, because my condition was not my back. When they did the MRI they came up with, you have arthritis. I have had no problem with arthritis. My problem was I have neuropathy from diabetes. . .

7.7. Some Patients Expressed Concerns Related to Advanced Images

Although several patients noted the benefits of advanced images, several reported concerns about the amount of radiation they were going to receive from an advanced image, or about the cumulative effect of receiving multiple scans. These participants

complained that their providers had either not addressed their concerns fully, or had not given them any information about this. Explained a male participant:

> I was here on the 6th also doing a CT-scan. I asked the nurse in charge, who also looked at my scan after it was done on my liver, whether the effects of CT-scans and all was cumulative. And she said yeah, but she didn't elaborate on it, and I was trying to kind of draw her out because every three months I'm having a CT-scan.

Another participant stated:

> Within the last three years, I've had it seems like quite a few tests here and I'm kind of curious how much of that radiation can you take before it starts dissecting your body. And I can't seem to get an answer.

All of the participants reported receiving their test results either by mail or via an electronic health record communication for patients. Explained a participant:

> I get my test results back and within 48 hours they're on my phone. And then they'll give you a call back, like, with maybe a weekend, that Monday or Tuesday they'll call you to confirm that they got the test results. And I do recommend [using the electronic health record for patients. It has] everything: your medicine, the department you can talk to; it has everything.

One focus group participant reported being frustrated by having to wait to get his test results:

> Did I follow up? No. I didn't try to check down the radiology department to find out what my MRI said, for instance. But I've had this situation before where test results were a bit vague coming back. Because if a patient is sitting around anxious, wondering what's going on, that would be a weak part I would say as far as that—gee we suspect this, we're going to do this. And then you wait and you're not quite sure whether it shows up in the mail, whether it's sent on to your doctor, your primary doctor. That's a bit of a vague point.

Focus group participants reported that their providers did not fully discuss the pros and cons of having an advanced image. Several patient participants reported that they did receive information on the risks associated with the test that had been ordered for them but that this information had been provided by a technician prior to conducting the test and not by the doctor. Explained one participant:

> Once you get to the department I found that the technicians are usually very—first of all, they were expecting you and that they're very receptive and . . . they explain to you as much as you need to know.

Another participant explained:

> . . . not the doctors, they don't say anything, they just schedule you down the line. But the technicians are very good at saying, particularly for an MRI, "you're gonna hear a loud noise, very loud in there. Not to move, you'll hear my voice." So they do walk you through it.

Such comments suggest that physicians are requesting multiple orders for advanced images and are not always fully explaining the need (or justification) for such images.

7.8. Conclusion

Overall, patients generally seem to like physicians' use of computers and see them as especially useful for providing information relevant to patients' health. However, most did were not aware that their physicians were using a DSS. Even without knowing DSSs were being used, patients had strong beliefs about advanced imaging, including the belief that they should have a say in decisionmaking about the use of images.

Patients acknowledge the physician's role in making decisions about images but also have a strong sense of their own involvement in the decisions. No patients complained of too much information about the use of advanced imaging. They all seemed to appreciate the information they had and some suggested they would have liked even more information about the reasons for the imaging and potential side effects, such as radiation exposure. Several wanted to better understand when radiation exposure with imaging was acceptable and when it would be best avoided. This suggests that patients want to be involved in decisions and appreciate the information they receive through computer-facilitated resources. Although focus groups did not explain the purpose of the DSS, we noted that if patients were to feel that a computer-associated program were denying orders that the patient had come to believe were appropriate, the patient might become resistant. Patients were appreciative of the advanced images they received and the computer programs that facilitated them. However, some reported they were not getting the "full story" about the pros and cons of these images, especially regarding serial images.

In summary, evidence from both physicians and patients suggests patients are not yet recognizing a specific impact from the MID, though they do recognize the many ways computers are being used to organize medical care and the delivery of services.

Section V: Six Statute Questions That Can Inform Future Recommendations About Decision Support

- Was the system for determining appropriateness in the demonstration acceptable for identifying appropriate versus inappropriate advanced imaging orders? (Chapter 8)
- Would exposing physicians to advanced imaging appropriateness criteria at the time of order affect the volume of utilization? (Chapter 9)
- Is it advisable to expand the use of appropriateness criteria for ordering advanced imaging to a broader population of Medicare beneficiaries? (Chapter 10)
- If expanded to a broader population of Medicare beneficiaries, should physicians who demonstrate that their ordering patterns are consistently appropriate be exempt from requirements to consult appropriateness criteria? (Chapter 11)
- To what extent is live feedback on the appropriateness of advanced imaging orders from a DSS better or worse than feedback reports to individual physicians or physician practices? (Chapter 12)
- In what ways can physicians be motivated—including financial incentives—to comply with ordering advanced imaging appropriately according to appropriateness criteria? (Chapter 13)

This page is intentionally blank.

8. Recommendations About the Acceptability of MID's DSS for Identifying Appropriate Versus Inappropriate Advanced Imaging Orders

This chapter presents recommendations about the acceptability of MID's DSS for distinguishing appropriate versus inappropriate advanced imaging orders, answering the statute question:

- Was the system for determining appropriateness in the demonstration acceptable for identifying appropriate versus inappropriate advanced imaging orders?

8.1. Summary Response to Statute Question

The use of the DSS for identifying appropriate versus inappropriate advanced imaging orders in the demonstration was not generally acceptable to clinicians in the form that was deployed. Furthermore, the system was not efficient, as it did not assign an appropriateness rating to more than one-half of the ordered advanced images.

8.2. Evidence

8.2.A. Clinicians Did Not Find the System Particularly Helpful

Clinicians who received and reviewed guidelines associated with their orders commented on the usefulness and validity of the guidelines in the focus groups and surveys. Table 8.1 shows the distribution of survey respondent ratings about the usefulness of guidelines. Overall, generalists, medical specialists, and surgical specialists were more likely to disagree than to agree with statements describing the guidelines as helpful.

Clinicians assigned a score from 1 to 9 according to the helpfulness of image-specific guidelines for each of the 12 MID images. Scores ranged from a value of 1, assigned to indicate "least useful" to a score of 9, assigned to indicate "most useful." The median rating for generalist and medical specialist focus group participants was 5, indicating these clinicians typically did not find the guidelines either helpful or unhelpful. In contrast, surgical specialist focus group participants assigned a median rating across MID guidelines of 2.75, indicating they did not typically find the guidelines very helpful. Across all three clinician focus group participant types, ratings were fairly similar across all the types of MID guidelines.

Table 8.1 Focus Group Survey Respondent Median Ratings about the Helpfulness of Specific Guidelines in Identifying the Usefulness of Images for Their Patients, by Specialty Type*

Specific MID Guidelines	Generalist N=27	Medical Specialist N=22	Surgical Specialist N=12
CT Abdomen	5.0	5.0	2.0
CT Pelvis	5.0	5.0	2.0
CT Abdomen and Pelvis	4.5	5.0	2.0
CT Brain	5.0	6.0	3.5
CT Lumbar Spine	5.0	5.0	1.0
CT Sinus	5.0	5.0	5.0
CT Thorax	5.0	5.0	2.0
MRI Brain	6.0	6.5	2.0
MRI Lumbar Spine	6.0	5.0	3.5
MRI Knee	6.0	4.0	4.0
MRI Shoulder	5.5	4.0	4.0
SPECT MPI	5.0	4.0	6.0

NOTE: Respondents assigned a score ranging from a value of 1 (least useful) to 9 (most useful) to the survey item, "For each image, how helpful is the guideline in identifying the usefulness of ordering this test for your patients?"

As discussed in Chapter 6, clinicians found the guidelines inaccessible, difficult to use in the context of their clinical practices, and not consistent with the kind of algorithmic clinical logic that they believe would provide more helpful clinical decision support.

A key component of successful appropriateness ratings is a precise and comprehensive listing of the reason for guidelines. This was not evident with MID. Furthermore, for guidelines to serve effectively as a decision support tool, they need to not only effectively synthesize available evidence, they also need to develop feasible strategies for linking the reason for the order as documented in available data sources to the guideline. As American health care systems are progressively shifting from paper to electronic health records and from ICD-9 to ICD-10 and other classification taxonomies, attention needs to be paid to valid methods for making these links.

Furthermore, guidelines need to be readily accessible to clinicians in the context of their clinical practice if they are to be used in real time to inform clinical decisionmaking. As discussed in Section 9.4, clinicians found guidelines with MID difficult to access and cumbersome to use when accessed.

8.2.B. More than Half of the Orders Placed Were Not Linked with a Guideline

As documented in Chapter 1, CMS selected the 12 imaging procedures for the demonstration (described above) because they are among the advanced imaging procedures that clinicians most commonly order for their Medicare patients. Analysis of the indications frequently associated with these procedures conducted prior to the demonstration by the implementation contractor,

The Lewin Group, indicated that between 53 percent and 89 percent of orders for these procedures should be successfully matched to a guideline and therefore be assigned an appropriateness rating by conveners' DSS systems during the demonstration. However, to encourage innovative models, CMS gave conveners the flexibility to determine how individual indications entered in a DSS would be linked to a specific guideline.

The actual percentage of orders receiving an appropriateness rating in both the baseline and intervention periods was much lower than anticipated (Table 2.4). In the baseline period, clinicians did not receive feedback about whether the order was associated with an available guideline, nor about the appropriateness of each order. We anticipated observing lower-than-expected rates of orders with appropriateness ratings during the baseline demonstration period because clinicians were becoming familiar with the ROE process during that period. We hypothesized that the percentage of orders that were linked to a guideline and rated would be higher during the intervention period after clinicians had mastered order entry. However, across image types, the percentage of rated images was 37.3 during the baseline period and 2.8 percentage points lower, at 34.5 percent, in the intervention period. Thus, even during the intervention period, the vast majority of orders were not rated. Orders without ratings were not subject to feedback about appropriateness. This resulted in the "dose" of the intervention being marked diluted, since almost two-thirds of the orders were not subjected to the main component of the intervention.

The differences we observed across conveners could be due to either structural characteristics of the convener's DSSs (particularly the way in which indications are selected) or differences in the specialty composition of individual conveners. If conveners recruited a large number of subspecialists who see patients with profiles that do not fit existing guidelines, we might see that the convener-level variability is really driven by differences in clinical specialties.

A major factor contributing to clinicians' limited enthusiasm for national specialty guidelines being considered useful relates to the large number of orders that were not covered by guidelines. Explained a Convener B generalist:

> I actually remember getting [a message stating that guidelines are not available] more than I remember getting an algorithm feedback with [an alternative order]. I get frustrated because I feel like, "okay, great, I just had to go [take my time to] scroll through—figure this out and now there's just no guideline to match this."

8.2.C. Clinicians Felt Constrained by Guidelines that Focused Only on Reasons for Orders

Many clinicians believed that guidelines could be more helpful to them if they were presented as part of a decision support algorithm rather than as support only for a single decision about the appropriateness of a particular procedure or image type. This is because patients often have many clinical challenges that have to be addressed simultaneously. Also, clinicians often make decisions about "what ifs," meaning they develop sequences considering how they should

behave if a clinical outcome differs from what was expected. A decision support algorithm could respond to a clinician's need to make a series of conditional decisions in sequence depending upon an iterative set of inputs.

Several MID focus group participants had been exposed to decision support algorithms to facilitate decisionmaking regarding a possible pulmonary embolism. Clinicians consistently contrasted how helpful they believed that algorithm to be compared with the singular focus of MID guidelines.

Furthermore, many clinician focus group participants indicated that they would be more comfortable with guidelines developed using a diverse, multidisciplinary team of raters, rather than one developed by the procedure-performing specialty. This is consistent with recommendations from the Institute of Medicine standing that guideline development should not be restricted to those from the specialty society associated with the procedure but should include other respected bodies less likely to be subject to potential or assumed bias (Field and Lohr, 1990).

8.2.D. Clinicians Were Not Consistently Comfortable with the Appropriateness Ratings Assigned by Specialty Societies

Consistent with recommendations by experts in the development and synthesis of ratings, many clinicians wanted even more transparency than was available about the appropriateness ratings. They wanted to better understand how ratings were assigned, especially when evidence was limited. This was particularly relevant because outcome studies that inform clinicians how use (or nonuse) of advanced imaging affects patient outcomes are sparse. Clinicians understand that expert judgment is often used to supplement existing trial and observational study data. Clinicians would like more consistent information across all guidelines regarding the quality of evidence used to generate the appropriateness guidelines, and more information about level agreement associated with appropriateness ratings. Clinicians also would like to have more input into the development of the appropriateness ratings by practicing clinicians, including both generalists and specialists.

The terms "appropriate," "uncertain," "inappropriate," and "not covered by guideline" may not convey the intended meaning well. With RAND's initial appropriateness work, the term "appropriate" was assigned when outcomes for a cohort of patients was expected to be more favorable with use of the procedure compared to without the procedure (Brook et al., 1986). In contrast, the term "inappropriate" was assigned when outcomes were expected to be less favorable without compared to with the use of the procedure. The term "uncertain" or "equivocal" was first applied to appropriateness ratings to represent one of two specific meanings. First it meant that the available evidence did not clearly distinguish better outcomes from worse ones for cohorts of patients with the use of the procedure in question, compared to without its use. The second meaning of "uncertain" or "equivocal" was that expert raters did not come to consensus with their rating, even after dropping outlier ratings. RAND concluded that

appropriateness ratings should be characterized by both (1) a metric representing whether outcomes associated with use of a procedure would be expected to result in better or worse outcomes for a cohort of patients with similar characteristics, and (2) a statement of the degree to which expert raters agreed with the recommended rating.

MID clinician focus group participants expressed interest in learning more about the meaning of the terms *appropriate* and *inappropriate*, what *uncertain* was intended to mean, and whether *uncertain* or *equivocal* were expected to convey a sense of less-than-optimal reason for ordering an advanced image.

8.2.E. Appropriateness Is Determined Differently Depending on DSS Design

Across the conveners, two broad categories of DSSs were implemented. The first system presented ordering clinicians and staff with a listing of discrete reasons for orders that could readily be mapped using DSS software to the guidelines that MID used for assigning appropriateness ratings. The second system presented clinicians with a series of clinical prompts that in aggregate allowed the DSS software to either query the ordering provider for additional data or to map to the guidelines that MID used for assigning appropriateness ratings. With the first system, clinicians were frustrated by long lists of drop down menus that were not organized according to clinical taxonomies familiar to the physicians or their staff. With the second system, clinicians were frustrated by the iterative queries that sometimes did not seem pertinent to the clinicians. With both systems, clinicians were frustrated by the extra time it took for them to scroll through long lists, particularly when the end result was often that their orders did not link to a guideline used by MID.

It is likely that the rates of linking to guidelines and the assignment of appropriateness ratings was influenced in some ways by the use of one or the other of these two systems. However, the large number of other variables that distinguished conveners did not allow us to systematically identify one critical feature of either system.

8.2.F. The Timeline for MID Implementation Was Too Short

All conveners noted that all phases of the MID timeline were too short to address the large number of challenges that needed to be addressed to successfully engage clinicians and staff with the planned demonstration, align existing and new workflow patterns required by MID, introduce clinicians and staff to the software issues and guidelines that were critical for MID's function, and to assure that the data and contractual requirements of the demonstration were met.

Across all phases of the timeline—from the period when conveners contractually agreed to participate in the demonstration, through the predemonstration window and the six-month baseline demonstration period, and finally to the 18-month intervention period—conveners reported inadequate time for setup, planning, pilot testing, implementation, internal evaluation, and change. Conveners' efforts to move forward rapidly were confounded by software

challenges beyond the control of the conveners and their practices, as well as escalating frustrations and disengagement by clinicians.

8.3. Recommendation

We need a more comprehensive set of guidelines that can cover a substantially greater proportion of advanced images ordered for Medicare beneficiaries.

- These guidelines need to more explicitly map the clinical characteristics that bedside clinicians are familiar with to the distinguishing features of the clinical guidelines.
- Alternative clinical and procedural options also should be made more explicit. They should include both radiologic and nonradiologic options.
- The guidelines need to consistently and explicitly document level of evidence as is standard for most published guidelines (Owens et al., 2010; Shaneyfelt and Centor, 2009; Grilli et al., 2000).
- Guidelines should be updated regularly (Shekelle et al., 2001).
- Guidelines should not be restricted to those from the specialty society associated with the procedure but should include other respected bodies less likely to be subject to potential or assumed bias (Field and Lohr, 1990).
- When guidelines are not available, more information should be available about the reasons why.
- Additional evidence should be explored about the best methods for incorporating levels of certainty into the guidelines and for making the rating system reflect that evidence.
- While an ordinal scale for measuring appropriateness has been used in many settings, this strategy should be further evaluated, especially when evidence is limited.
- More research is needed on DSS design to better understand circumstances when DSS feedback is most needed: while a clinician is deciding whether to order an advanced image, or after the decision to order the image is made.
- Guidelines need to be more explicit about how appropriateness might vary for initial compared with serial images. The appropriateness of serial images when signs and symptoms are escalating or resolving need to be distinguished from the timing of serial images when patients are stable. Guidelines need to be more explicit about periodicity.

9. Recommendations About Volume of Utilization in Response to Physician Exposure to Advanced Imaging Appropriateness Criteria at the Time of Orders

This chapter presents recommendations based upon multiple chapters presented earlier in this report to address the statute question:

- Would exposing physicians to advanced imaging appropriateness criteria at the time of order affect the volume of utilization?

9.1. Summary Response to Statute Question

Empirically, within MID, exposing physicians to advanced imaging appropriateness criteria at the time of order did not have a substantial impact on volume.

9.2. Evidence

9.2.A. Evidence from Claims

In Chapter 4, we observe that providers who receive higher proportions of inappropriate feedback about advanced image orders are no more likely to change their future volume of orders than are providers with higher proportions of appropriate feedback. In Chapter 5, we observe a drop in utilization that is significant from the predemonstration to the demonstration period for all but one of the conveners. However, this drop does not differ significantly for demonstration versus comparison sites for five of seven conveners. For the two conveners with statistically significant reductions in the number of advanced images utilized by the demonstration compared with the matched control providers, the estimated difference is relatively small, implying an additional one MID image per 100 Medicare beneficiaries reduction within demonstration versus control sites. Given that even these two significant results are sensitive to whether images are included in months when no beneficiaries are found for a particular provider, we consider the evidence for these demonstration effects to be weaker than the confidence intervals imply on their own. In aggregate, our analyses provides no evidence that appropriateness feedback leads to anything beyond a small reduction—if any reduction at all—in MID image volume.

9.2.B. Evidence from DSS Analyses

Nevertheless, Chapter 2 identifies important evidence that clinicians do respond to some components of the intervention, particularly to the feedback of alternative orders as they make decisions about changing and cancelling orders. These data suggest the conceptual model for a

DSS is acceptable to clinicians. However, the lack of major changes overall, as documented in Chapters 2 through 5, highlight challenges associated with implementation of MID, as have been well documented in earlier chapters of this report and in RAND's Focus Group Report to CMS (Kahn et al., 2013). If the clinicians were more engaged, if the user interface were more acceptable to clinicians, if the electronic health record were linked with DSS, if changing or cancelling an order in response to DSS feedback were not so time consuming, or if clinicians were able to better understand how to link their reasons for orders with the guidelines, then it is possible (or even likely) that volume could change with a DSS. However, none of these counterfactuals has been changed since the initiation of MID. Beyond that, our empirical analyses show no evidence for volume changes related to DSS implementation.

9.2.C. Evidence from Focus Group Analyses

While clinicians supported the key ideas behind the demonstration, particularly the need to reduce the number of unnecessarily or inappropriately ordered advanced imaging, they did not feel that the design of this demonstration project was conducive to producing such changes. One key reason for their negative assessment of the demonstration concerned the way the DSS worked. Clinicians had hoped for clinically meaningful data to be made available to them in real time, but they reported that this type of clinical feedback was not provided or that they received the feedback too late.

At most sites, these problems were compounded by difficulties in cancelling or changing an order, in the context of placing an advanced imaging order while using the DSS. Within most practices, clinicians placed orders within their EHR, which involved a different workflow than the one required to use DSS. Thus, any changes in ordering that a clinician might want to make based upon DSS feedback were confounded by time-consuming switches between DSS and the EHR.

When focus group participants were asked about the potential effect the appropriateness feedback could have on their advanced imaging utilization, many either did not provide clear answers or stated that they needed to see the actual numbers and could not base their answers on the limited feedback they received so far. Nonetheless, some Convener C and E specialists reported concerns that rising health care costs represent a complex problem that requires changes at multiple levels—not just in a reduction in advanced images. A Convener E specialist stated that, in many situations, s/he was concerned about the repercussions involved in *not* ordering an advanced image:

> I guarantee you, 50 percent of the MRIs that I order, the radiologist will comment, saying that there is an insignificant spot in the liver and will automatically recommend a repeat MRI in six months. . . . I will go out on a limb if I do not order that MRI and put myself at medical legal risk if I truly believe it's a potential lesion. But I think the problem is huge. I don't think the problem is just the person who orders the film. I think the problem is the person who orders the film, the person who reads the film, and the person getting the film.. . . I'd like to be more optimistic, but it's just a very complex question.

Focus group participants also felt that significant changes in order utilization, appropriateness, and overall cost would require dramatic shifts in culture, such as moving away from "defensive medicine," the strategy of recommending an image order that is not necessarily indicated for the patient but serves to protect the physician against the patient as potential plaintiff. Focus group participants did not believe this demonstration could have addressed all of these topics. In the words of one Convener C specialist:

> I think the evidence would suggest that we probably can do with less, but in each specific instance, you're ordering the test because you think it's necessary. So it's really a change of culture and how we practice medicine and that's not something that I think this system is going to easily be able to do. . . . When you have to discuss it with an insurance company, which is still frustrating, you're at least talking to somebody. You can have a back-and-forth discussion. In an electronic system, which has its advantages, it also has a disadvantage that you're not really talking to someone who you can have a dialogue with, and that's where a lot of the frustration comes in, is no ability for the system to give and take as much as it wants us to.

Participants suggested that the introduction of DSSs alone is unlikely to affect health care culture. Doing so would require a comprehensive reform that was grounded in data-driven approaches that provide clinicians with more detailed and complete data on utilization and appropriateness, more thorough education about alternative orders, and strategies for how to interact with patients who insist on ordering images deemed inappropriate by DSS.

Moreover, some specialists felt that it is not only clinicians' culture, but also patients' culture and values that need to be changed before the number of inappropriately ordered images can be reduced. As a Convener C specialist stated,

> This is not just the physicians, this is the cultural thing of patients and physicians, that entire group is not willing to wait, the pressure is on everyone to do it earlier. . . . I think it's not so much appropriateness, but . . . the redundant testing that patients who do get referred into other institutions who've already had tests elsewhere. . . . As electronic med records and transfer of information improves, that will also improve, but we aren't there yet.

Guidelines themselves cannot be used in isolation to distinguish appropriate and inappropriate image orders. Such discrimination depends upon precisely linking the ordering clinician's reason for the image order with the relevant guideline. This type of precise linkage requires clinicians to enter into the ROE system details about the reasons for the order. This entry has to be facilitated by a clinician-friendly user interface that allows the clinician to reliably and validly document the reason for an order within a brief time window. This order entry function has to incorporate the fact that a clinicians' conceptualization of how an advanced image might help them manage their patient's problem may not match precisely with prespecified reasons for orders that map to existing guidelines.

Furthermore, within MID, the lack of integration between the EHR and the DSS created a challenge for clinicians efficiently documenting the reasons for orders in a manner that would seamlessly link with the relevant guideline. Simultaneously, the guidelines need to specify the

clinical detail necessary to discriminate more-appropriate ratings for each image type, compared with less-appropriate ones.

Some clinicians have suggested that feedback about advanced image orders would be more effective if provided as they formulate the reasons for order, rather than after they place the order (even if the feedback is received immediately after placing the order). Many clinicians expressed frustration that their appropriateness feedback reflected an inadequate report card grade that was handed to them by a poorly specified computer program that did not understand clinical practice and did not know their patient. Instead, they wanted a DSS program that would guide them through the decisionmaking process so they would place an advanced image order only when probabilities for a good patient outcome were improved with use of the advanced image, compared with no use. They contrasted this prospective system of receiving decision support prior to having decided to place an order with their experience of MID's DSS, which many described as generating a feeling of having received "a bad report card" after working hard to enter the data but having learned nothing about how to improve care for their patient.

Clinicians indicated they would be willing to provide clinical data to the order entry system provided the information would be useful in helping them decide whether to order advanced imaging and, in that event, it advised which specific image would be most likely to improve outcomes of interest. Clinicians also wanted reassurance that the next generation of DSS would make use of clinically relevant data that they had already entered into their EHR.

9.3. Recommendation

While some aspects of the DSS program seem promising, they did not translate into observable changes in utilization for MID in relation to comparison sites. In some sense, this is not surprising, because a DSS should help clinicians perform their work more efficiently and effectively. It should not hold up their time, taking them away from other important patient care responsibilities. A DSS should facilitate easy order entry, specification of the reason for order, receipt of appropriate orders, and awareness of alternative interventions (radiologic or nonradiologic) that could improve patients presenting with the problem the clinician is assessing.

Based upon these findings, we offer the following recommendations:

- Addressing each of the identified challenges associated with the MID implementation appears to be a prerequisite before DSS utilization will lead to a reduction in the use of inappropriate advanced image utilization.
- More research is needed about how to engage clinicians with the guideline development process. These efforts should study how clinician and staff users can be involved in the development and use of guidelines. Even when local systems are used to develop or adapt guidelines, local users should be engaged, so that their clinical inputs are incorporated into guideline development and testing.
- More effort is needed to design easy-to-access guidelines, ratings, and alternatives.
- Further software linkages need to be implemented between clinical data and clinical details housed within electronic health records and administrative data.

10. Recommendations About the Advisability of Expanding the Use of Appropriateness Criteria for Ordering Advancing Imaging to a Broader Population of Medicare Beneficiaries

This chapter presents recommendations based upon multiple chapters presented earlier in this report to address the statute question:

- Is it advisable to expand the use of appropriateness criteria for ordering advanced imaging to a broader population of Medicare beneficiaries?

10.1. Summary Response to Statute Question

As implemented, MID has involved a broad population of practices, providers, and Medicare beneficiaries. The results of our analyses suggested that it is not advisable to expand the use of appropriateness criteria for ordering advanced imaging to a broader population of Medicare beneficiaries. Our analyses suggest that the demonstration is not ready to be scaled up and that several issues will need to be resolved before scale-up will be viable.

First are the issues raised in Chapters 8 and 9 concerning the acceptability of MID's DSS for distinguishing appropriate versus inappropriate advanced orders, and ultimately for reducing the number of inappropriate orders. Our analysis found that the demonstration was not successful in either respect, as explained in the previous chapters.

However, even if the guidelines were improved and the system showed a higher success rate in reducing the number of inappropriate orders, there would still be additional implementation issues that require attention before any scale-up could be considered. The implementation period was too short to ensure that the DSSs were up and running and to address problems in linking to EHRs. Second, there are important issues around workflow with DSSs and how clinicians feel about DSSs compared with prior authorization, a yet-to-be-tried system of prospective decision support for clinicians at the time they are deciding to order an advanced image, or other evidence-based clinical interventions.

A third issue concerns the tension between wanting consistent standards across all appropriateness guidelines and achieving buy-in from physicians by filling in guideline gaps with local best practices.

10.2. Evidence

The MID has been implemented across a broad population of beneficiaries, including seven conveners spanning multiple regions of the country, states, and delivery systems. In this sense,

the demonstration's roll-out applied to a broad population of Medicare beneficiaries from the inception of the demonstration.

The evaluation of MID has recognized some gains in the proportion of images that are appropriate (see Chapter 2), and no increase in rates of use of advanced imaging since the demonstration began (see Chapters 4 and 5). Nevertheless, the evaluation also revealed substantial frustration, lack of satisfaction, and missed opportunities that if addressed, could lead to substantial improvements in the rates of appropriate advanced imaging procedures.

These issues should be addressed before a major scale-up is implemented.

10.2.A. Clinicians Need to Be Better Engaged in What a DSS Can and Cannot Do at Present, While Preserving Enthusiasm for How Its Effectiveness Will Improve with Time

Addressing challenges to clinicians' efficient use of DSSs should be acknowledged and resolutions explored with substantial inputs from clinician users. These challenges should be resolved prior to expanding the use of appropriateness criteria for advanced imaging to a broader population of Medicare beneficiaries. For example, access to guidelines and evidence supporting them needs to be more readily available for ordering clinicians. The user interface needs to be improved so reasons for orders can be efficiently entered into order entry. Guideline transparency should be greater regarding level of evidence and level of agreement. This is consistent with the approach advocated by the National Guidelines Clearinghouse (U.S. Department of Health and Human Services, undated). Linkages between DSSs and EHRs should be improved so that the clinical detail entered into one system can populate relevant fields in the other. Substantial effort still needs to be developed to properly allow this data transfer to occur. Furthermore, the logic for and rules associated with transferring these data also needs to be pursued. Clinician's use of DSSs compared with prior authorization will need to be further explored as efficiencies in both systems are enhanced.

10.2.B. Additional Thought Needs to Be Given to the Time Required for a Successful Implementation Period

This time will vary according to the degree of cohesion between the clinical leaders and those who order advanced imaging, as well as the alignment between the clinical team and the information technology and administrative leaders. Clinical settings with infrastructure in place for implementing innovation are more likely to require shorter implementation periods, as are those where clinicians and staff already participate in a regular program for learning about new innovations in health care delivery.

10.2.C. Strategies for Integrating National and Local Guidelines Standards Should Be Developed

Using national guidelines supported by a vigorous evaluation and review process provides a framework for consistency across the heterogeneous providers and practices throughout the nation. However, tensions between national guidelines and local practices are prevalent and strategies for addressing these tensions should be addressed prior to scale-up. Development of a protocol for populating DSS with guidelines that consider both national and local guideline inputs is likely to enhance buy-in by a broader coalition of clinicians while also providing the opportunity to fill in guideline gaps with local evidence.

10.2.D. Optimal Strategies for Engaging Patients in Knowledge About and Decisionmaking Regarding Advanced Image Use Should Be Explored

While Chapter 7 indicated that patients generally are not aware that clinicians' ordering of advanced images may be influenced by a DSS, patients themselves are eager to learn more about when they should have initial and subsequent advanced images and why. Furthermore, patients want more information about what to expect during the image ordering procedure, what impact the results are likely to have on their health, and how they can remain current with projections about when the radiation exposure and the fiscal costs associated with advanced imaging should be a concern for them.

10.3. Recommendation

As implemented, MID already has involved a broad population of practices, providers, and Medicare beneficiaries. Prior to expanding the use of appropriateness criteria for ordering advanced imaging to a broader population of Medicare beneficiaries, strategies should be explored for addressing already identified challenges to the use of appropriateness criteria.

In particular, strategies for further engaging clinicians and staff, improving the quality of guidelines, and improving ordering workflow (including the DSS user interface) should be addressed.

10.3.A. Expertise Within and Across Multiple Disciplines Will Be Required to Address the Current Challenges to Effective DSS

- While teams of discipline-specific expertise should explore options for addressing challenges, cross-disciplinary teams will also be essential to align priorities for clinical, information technology, administration, and patient stakeholders. Clinical inputs must be responsive to the workflow requirements of the set of generalist and specialist physicians who traditionally have been responsible for advanced image ordering, as well as to the growing set of ancillary and support staff who currently implement many tasks previously performed only by clinicians. To assure the new generation of physicians,

advanced-practice nurses and physician's assistants, nurses, and other staff are familiar with DSSs, their function should be further integrated into training curricula for these providers.

- Informaticists working to develop EHRs and clinically meaningful decision support tools should align with those developing DSSs for advanced imaging. Radiologists and clinicians should consider how workflows pertinent to order entry and documenting reasons for orders can be aligned across departments. Recommendations for changes to ordered images and reasons for changes should be apparent to both clinicians and radiologists.
- Protocols for assigning reasons for ordering advanced images should be known to those involved with ordering. These protocols should be consistent with both national and local guidelines.
- Discrepancies noted between national and local guidelines and between radiology and other specialty specific guidelines should be explored. Reasons should be specified so that relevant stakeholders understand reasons for ongoing discrepancies.
- When a DSS is used, it should be actively managed such that clinicians associated with orders that are not linked to guidelines or that are frequently inappropriate should engage in discussions about the reasons for this. These reasons should be vetted with other clinicians, radiologists and administration. Administration should consider discussing these issues with payers to further discriminate valid versus other reasons for discrepancies.

10.3.B. Specific Efforts Should Be Provided to Support Clinicians as They Undergo Major Changes in Health Care Delivery

The introduction of DSSs at the same time clinicians were introduced to EHRs has put substantial strain on those who already practice long hours caring for patients. While the dissemination of electronic information technology into health care structure and service use may only be at the beginning of its trajectory, the impact on clinicians has already been enormous. Providing adequate training, attending to how workflow could be affected, and supporting clinicians as their productivity is challenged can make an enormous difference in the degree to which clinicians buy into the process.

- Problems should be identified and addressed using a multidisciplinary team as needed to ensure that the clinical, administrative, patient engagement, fiscal, and information technology issues are aligned.

10.3.C. Identify the Specific DSS Challenges Associated with Subpopulations

- Given the complexity of addressing the many challenges to DSSs that have been recognized, it may not be possible to address all challenges at once. Instead, identifying subpopulations that could benefit from specific interventions would be helpful.
- RAND's analysis suggests the following provider types are worthy of additional study:
  - Providers who order very high volumes of advanced imaging especially when they order the same test for a limited number of reasons. If they are not engaged with DSS or if they disagree with the guidelines, their volume could have a significant impact on a practice's appropriateness ratings and on the way others practice. This is especially true if they function as local opinion leaders. Strategies should be developed to identify these clinicians and engage them in learning more about how DSSs could support their efforts while also learning from them how DSSs could be more helpful.
  - Providers who consistently order advanced images that are not associated with guidelines. These providers are likely to have substantial inputs into how the guidelines could be restructured to better engage them.
  - Providers who order low volumes of multiple types of advanced images (such as generalist clinicians) who may need a different type of feedback report than those who order one image type with high volume.
  - Providers who have substantial experience using ERHs and other DSSs versus those who are novices at using even an electronic order entry system.
- Stratifying scale-up efforts to focus on guidelines that are well specified, well supported with evidence, and consistently agreed upon would facilitate an evaluation that could be more nuanced in determining which aspects of the DSS intervention make a difference.
- Working with guideline developers to ensure that guideline ratings are available for both first-time and follow-up orders, with each set being supported with evidence will provide clinicians a better understanding of when imaging can improve their patient's outcomes.
- Given the many factors that have been identified as influencing the clinician's and practices' experiences with DSSs, scaling up with a more focused approach could allow the process to reach an earlier tipping point of acceptability.

This page is intentionally blank.

11. Recommendations About the Advisability of Allowing High-Performing Physicians to Be Exempt from Requirements to Consult Appropriateness Criteria

This chapter presents recommendations to address the statute question:

- If expanded to a broader population of Medicare beneficiaries, should physicians who demonstrate that their ordering patterns are consistently appropriate be exempt from requirements to consult appropriateness criteria?

11.1. Summary Response to Statute Question

RAND has conducted empirical analyses of the stability of rates of appropriateness to advise whether certain types of physicians should be exempt from a potentially broader appropriateness criteria program and whether this exemption should be affected by image type and/or by patient characteristics. Using a decision-theoretic approach, it is feasible to provide a model to address the question of what types of ordering histories should exempt providers from using DSSs and for how long. However, the meaningful application of such an exemption model is predicated by the assumption that a DSS is broadly useful in the setting where the exemption criteria would be considered. In fact, this evaluation has not demonstrated that DSSs is broadly useful for advanced imaging, as implemented with MID. Thus, this chapter presents an approach to decisionmaking about exemption from DSS use that could be of value in a setting where DSSs for advanced imaging were useful.

11.2. Evidence

For providers who consistently have high rates of appropriate orders, the information provided by a DSS may not be worth the time it takes to navigate the system. Hence, we consider methods for determining whether it is appropriate to exempt certain providers from DSS use—and if so, for how long. Note that any discussion of exemption from DSS use is predicated by the assumption that a DSS is broadly useful. If a DSS is not of sufficient utility to provide benefits to a substantial proportion of the exposed providers, it should not be rolled out, regardless of whether providers may be exempted.

We propose a decision-theoretic approach to address this question of what types of ordering histories should exempt providers from DSS use, and for how long. That is to say, we quantify the relative importance or "loss" associated with making particular types of "mistakes." In this context, we assume there is a desired *appropriateness rate* for each provider, by which we mean the expected proportion of appropriate orders out of all rated orders during a given time period.

For example, if a provider only has two rated orders in a given period of time, we will only be able to estimate the appropriateness rate for that provider with a great deal of uncertainty since the appropriateness rate reflects the expected proportion of appropriate orders, not merely the observed proportion of appropriate orders.)

Overall, when considering potential exemptions from using DSS, there are two mistakes that we can make: exempting a low-appropriateness provider (i.e., a provider whose current appropriateness rate is below the desired level) and failing to exempt a high-appropriateness provider. Having quantified the "loss" associated with these two types of possible mistakes we may—after specifying a statistical model for the appropriateness of each provider's orders—derive optimal exemption decisions given each provider's ordering history.[12] The actual thresholds for exemption—and the length of time for which an individual can be exempted—depend on several key parameters that we will discuss.

When quantifying our relative "losses" (i.e., the relative costs associated with making the two types of mistakes outlined above), it is important to recognize that different stakeholders may have differing opinions on the various tradeoffs involved. For example, the loss could be elicited from the perspectives of hospital administrators, insurers, physicians, patients, or others. Although there may be broad qualitative agreement on some goals (increasing hospital efficiency, decreasing patients' radiation exposure, increasing the utility of rendered images) the weight that is given to such domains may differ substantially by one's perspective. For instance, a hospital administrator may place higher emphasis on a provider's ability to see a large number of patients in a day and lower emphasis on reducing overall spending on advanced images than CMS would. While differences in opinion can affect the specification of the loss function, the proposed method can be carried out with any entity's perspective built in.

Throughout the discussion, we refer to *exemption* from DSS use. We note that this term may have several meanings depending on the clinical environment. For example, in some hospital systems, it may be the case that some amount of information on the reason for order is required of all imaging orders. In such situations, "exemption" would refer to exemption from the additional data entry that would be required to produce a DSS rating. In other situations, it could mean that providers are not required to use any prespecified electronic order entry system to enter the reason for order; instead they could use a free-form system as clinicians have used for decades.

[12] Our statistical model does not include explanatory variables, and each provider's appropriateness rate is estimated independently of all other providers. Jointly estimating the appropriateness rates of all providers would be expected to improve the overall accuracy of our provider appropriateness estimates, but low volume providers would likely be exempted solely based on the performance of others. In our approach, providers' records must speak for themselves and are not impacted by their peers' performance.

11.2.A. A Model for Exemption from DSS

Our approach to setting cutoffs for exemption from DSS relies on several key assumptions that are summarized here, then discussed in greater length:

1. We are able to specify a target appropriateness rate for each provider. This may be different for different provider groups.
2. The probability of an order being appropriate is constant within periods.
3. There is a small, known probability of change in appropriateness rates between periods. This probability can be tuned to yield decision rules with face validity.
4. If there is a period-to-period change in appropriateness rates for a provider, we assume that the new appropriateness rate is drawn independently of the previous value.
5. The appropriateness ratings of orders are independent of one another (given the true period-specific appropriateness rate).
6. The relative loss associated with the two types of errors (exempting low-appropriateness providers or failing to exempt high-appropriateness providers) can be quantified and the relative loss depends only on the provider's ordering patterns relative to the target

First, we assume that administrators can set target proportions of appropriate orders that each physician would be expected to meet. For example, an administrator might wish to enforce the goal of having—among orders that are rated—75 percent appropriate (rather than inappropriate or equivocal) orders for one group of providers. This target may vary by assorted factors, such as specialty and place of service, but we assume that, for each provider, administrators are able to specify such a target.

Second, we assume that if a physician's true appropriateness rate is greater than the target, there is a net "loss" associated with subjecting that physician to DSS use. For example, if a physician's orders are highly appropriate, the time associated with navigating the DSS system could be worth more than the information he or she is receiving. On the other hand, for a low-appropriateness provider, we assume there is a net loss if he or she is not exposed to a DSS. For such providers, we assume that the information they would receive from a DSS is worth more than the time that it takes to get the DSS feedback. For the model described below, we assume that the loss is constant for all physicians whose true appropriateness is above or below the cutoff. Thus, for the 75-percent appropriateness goal, the loss associated with failing to require DSS use is assumed to be the same for providers who have either 60 percent or 70 percent appropriate orders.

Further, although we may care more about getting the decisions right for high-volume providers, we assume that the relative trade-offs do not depend on volume. For example, if one were able to denominate the two possible types of mistakes in dollars (i.e., exempting a low-appropriateness provider and failing to exempt a high-appropriateness provider), it may be the case that they "cost" the decisionmaker $2,000 and $1,000, respectively, for a high-volume provider. For a low-volume provider, these costs may be $200 and $100. Even though the

magnitudes are different, the relative sizes of the costs associated with the two types of mistakes are equivalent. We feel that this assumption will generally be reasonable; if not, the loss could be specified to depend on provider volume.

We assume that there is zero net loss associated with exempting high-appropriateness physicians from DSS use or with exposing low-appropriateness physicians to it. If these assumptions on the loss specification hold, the administrator needs only to provide a single, relative loss value of the form "the loss associated with exempting low-appropriateness physicians is X times as large as failing to exempt high-appropriateness physicians." For example, one might say that it is twice as bad to exempt low-appropriateness providers as it is not to exempt high-appropriateness providers; this would complete the specification of the trade-offs between our two types of errors.

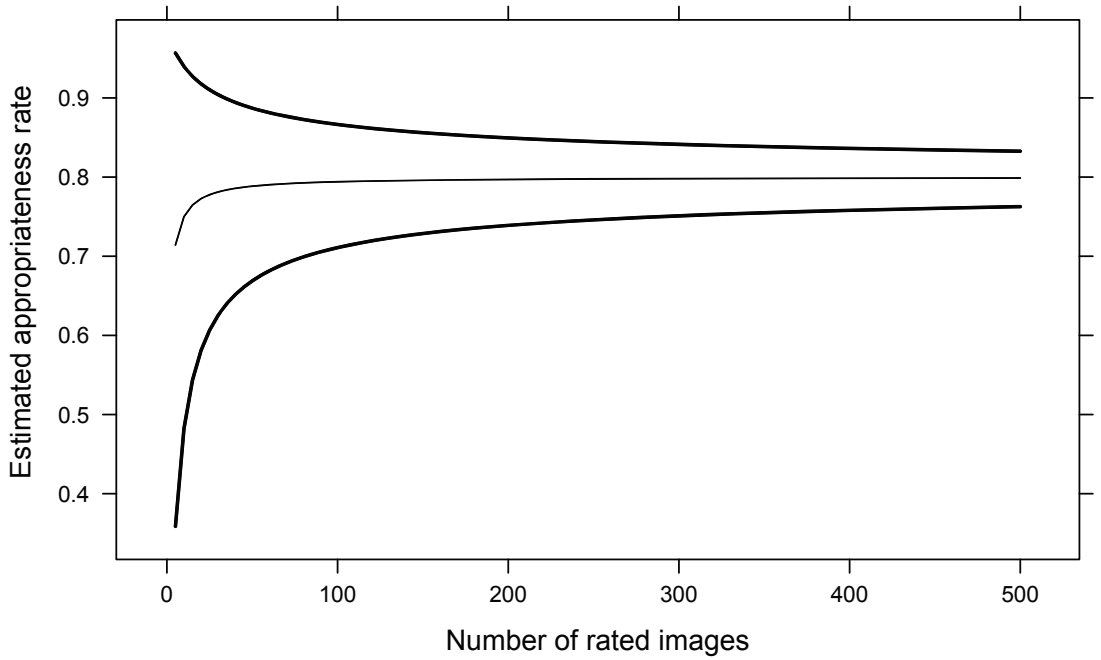
Next, we assume that providers' expected appropriateness levels are likely to stay fixed over time but that there is a small probability of change between the periods (such as months) for which exemption decisions are made. We denote this probability of a period-to-period change in the appropriateness rate as p_c. This parameter is key for the determining the length of time for which high-appropriateness physicians may be exempted. Hence, we will assume that a fixed, known value of p_c will be chosen by administrators in order to achieve qualitatively desirable lengths of exemption, though it would in principle be possible to learn p_c from the data. Including this probability of change in appropriateness rates is necessary in this modeling approach. Without it, a provider who was exempted for one period would be exempted forever. This discussion of changes in appropriateness rates assumes that the providers have reached a steady state in their appropriateness rates, such that initial learning from the start of DSS use has already occurred. For this reason, we do not recommend turning on the exemption criteria immediately after a DSS is rolled out in a particular clinical environment.

Further, we assume that the appropriateness ratings of each provider's orders are independent of each other, conditional on each provider's current appropriateness rate. By this we mean that, although there may be substantial provider-to-provider heterogeneity in appropriateness levels, knowing the appropriateness of one order does not give us information about the appropriateness of any other order, beyond what it tells us about the provider's true *appropriateness rate*. This assumption is likely to be imperfect in the case of repeat image orders for individual patients (where the indications and therefore the appropriateness may be the same), but we expect that violations of this assumption will have a minor impact on exemption decisions for most practice patterns.

To finish specification of the model, we next assume that, *a priori*, providers' true appropriateness rates are drawn from a uniform distribution between 0 percent and 100 percent. In the case of a change of appropriateness for a provider (the probability of which is described by p_c), we assume that the new draw is taken independently of the previous value. Hence, if there is a change in appropriateness rates, we do not assume ahead of time that it would be more likely to be a small change than a large change.

The previous paragraph describes an example of a Bayesian prior distribution. If little information (i.e., small number of rated orders) is available, this means that we will infer the appropriateness rate with little precision. As more information becomes available, our estimate of the underlying appropriateness rate moves closer to the empirical average, and the uncertainty in the estimate decreases. An example of this is given in Figure 11.1. For a number of rated images between 5 and 500, the figure gives the estimated (thin black curve) appropriateness rate, along with 95 percent intervals, assuming that exactly 80 percent of the observed orders are rated as appropriate. Accounting for the sample size–dependence of the uncertainty around these estimates is vital for this application; making the optimal exemption decision depends on the level of confidence we have that the provider's appropriateness rate is above the desired level, not on the estimate of the appropriateness rate.

Figure 11.1. Bayesian Estimate of Appropriateness Rate for Samples of Different Sizes with 80-Percent Appropriate Orders



NOTE: The thin curve represents the estimate and the thick curves denote the ends of 95-percent intervals.

A consequence of this method's dependence on uncertainty of estimates rather than merely the point estimates is that there is no minimum volume of orders needed to apply the method. If a provider has a small number of orders and all are rated appropriate, there will still be enough uncertainty in the rate that such providers would not be exempted under typical specifications of the loss values.

A related issue is the potential for providers to enter information such that the DSS returns a rating of "not covered" for all but a small fraction of the orders that a provider knows will be

rated appropriate (from previous interactions with a DSS) in an effort to game the exemption system. Our recommendation in this regard is that providers need to clear a "gate" in terms of the percentage of their orders that are rated in order to be eligible for exemption. The most straightforward approach would be to require that providers had at least a certain percentage of rated orders out of the last, say, 50 orders to even be considered for DSS exemption. An alternative would be to use not merely the number of rated orders in the denominator used to estimate appropriateness rates, but the total number of orders, including those that are not rated. We recommend that this latter approach only be implemented for specialties for which the DSS has very good coverage—otherwise, providers with a substantial portion of their orders that genuinely are not covered by a DSS would be penalized.

The above paragraphs describe all of the assumptions that are required for our model. Under these assumptions, we are then able to derive optimal exemption decisions for each provider on a period-by-period (e.g., monthly) basis.

11.2B. Examples of Applications of the Proposed Decision Rule

We now apply our proposed exemption rules to several hypothetical provider order histories. We assume the period-to-period probability of a change in appropriateness rates is p_c=5 percent and assume that the target appropriateness level is τ=75 percent. We also assume that the loss associated with exempting a low-appropriateness provider is twice as high as failing to exempt a high-appropriateness provider. First, consider a provider who has consistently high appropriateness ratings. We assume that provider X has a 12-month history with 15 rated image orders each month. We simulate the appropriateness of orders such that each image has an 85-percent probability of being rated appropriate. A realization of such monthly appropriateness counts (out of 15 rated images) is

[13,12,14,12,13,15,13,12,11,13,7,10].

The empirical estimate of provider X's appropriateness rate for this period — at around 81 percent — is a little below the true appropriateness rate of 85 percent. Even so, applying the proposed decision rule, we see that provider X would be exempted for the next 10 months.

Now consider provider Y, who has the same 12-month order history, except with a different number of appropriate orders out of the last month's 15 rated orders. If provider Y's 12th month saw three of 15 appropriate, provider Y would not be exempted in the following month. If four of 15 in the last month were appropriate, provider Y would be exempted, though only for two months. If all 15 were appropriate, provider Y would be exempted for 11 months, or one more than for the original history.

As it turns out, the longest possible exemption period with p_c=5 percent is 11 months. With a smaller p_c=2.5 percent, the longest possible exemption period is 23 months. With a larger p_c=10 percent, the longest possible exemption period is five months. (Note that all of these numbers depend on the appropriateness goal τ as well as the relative loss for the two types of

possible mistakes equal to two. Larger values of τ or this relative loss would result in shorter exemption periods.)

11.3. Recommendation

11.3.A. Identify an Effective DSS Application Prior to Considering Whether Some Providers Should Be Exempted for Periods of Time

For this exemption scheme to be useful, we highlight several broad considerations already mentioned. First, the methodology assumes that the DSS is, in general, a tool that provides valuable information for providers. If the information the DSS provides is incommensurate to its burden for most providers, having exemption criteria in place should not substantially change the decision whether to roll it out in clinical practice in the first place. Hence, the utility of the DSS as a whole should be evaluated before considering whether some providers should be exempted for periods of time.

In the context of this evaluation, and based upon the empirical evidence obtained from quantitative analyses using DSS data (Chapters 2, 3, and 4) as well as from claims data (Chapter 5), and from quantitative analyses of focus groups with clinicians and staff (Chapter 6), at this time we cannot be certain that the information providers receive from using DSS justifies the cost in time associated with its use. On the other hand, as noted in Chapters 8, 9, and 10, we anticipate that DSSs could be improved so that clinicians using an improved DSS to get evidence-based guideline data at the time of order entry could substantially improve ordering practices. Once a DSS is recognized as systematically effective for its users, then the exemption model presented here is likely to have value for users.

11.3.B. Achieve Consensus on the Relative Losses Associated with Exemptions Using Input from a Broad Set of Stakeholders

The most subjective portion of the proposed exemption process is the specification of the loss associated with exempting a low-appropriateness provider relative to failing to exempt a high-appropriateness provider. As mentioned above, patients, hospital administrators, providers, and insurers may have very different views on these trade-offs. Insurers and patients may attach a relatively higher loss to exemption of low-appropriateness providers from DSS out of concerns related to spending and radiation exposure, whereas providers and hospital administrators may attach a lower loss to such mistakes because of greater concern for provider efficiency and financial incentives in fee-for-service systems. No single perspective is inherently correct, but we emphasize that the decisions that are optimal from one perspective may be less desirable from another. To the extent possible, we recommend that a consensus on the relative losses be reached using input from all interested parties. Such an exercise would require an intimate understanding of both the costs and benefits of DSSs. Given that these costs and benefits can depend on

provider specialty, clinical location, and other factors, an optimized exemption system would likely not be the same for all providers.

11.3C. Consideration Various Types of Exemptions

A potential extension of the proposed system would be to track the types of orders that are frequently made by each provider. (This would only apply in clinical environments where providers are required to enter minimal reason for order information, even if they are exempted from DSS use.) If, for example, an exempted provider were to order an image on a body part that is unusual compared with their usual practice pattern, DSS feedback could be provided for that single order. The justification for such a suspension is that the exemption was "earned" on orders that are unlike this new one, so it is not clear that the exemption should be applied here as well.

12. Recommendations About the Value of Live Feedback on the Appropriateness of Advanced Imaging Orders from a Decision Support System Compared with Feedback Reports to Individual Physicians or Physician Practices

This chapter presents recommendations to address the statute question:

- To what extent is live feedback on the appropriateness of advanced imaging orders from a decision support system better or worse than feedback reports to individual physicians or physician practices?

12.1. Summary Response to Statute Question

Since the limited number of orders submitted by many clinicians has limited the extent of dissemination of aggregate practice- and person-level feedback that has been distributed, RAND has not been able to conduct empirical analyses of the relative impact of aggregate feedback reports to individual physicians or physician practices.

12.2. Evidence

The design for the demonstration was to first expose clinicians to the baseline intervention period for six months, then expose them to 18 months of concurrent feedback about the appropriateness of the order at the time the order was placed. Additionally, at periodic intervals during the demonstration, individual conveners, practices, and physicians were to receive aggregate feedback reports covering an ordering window of several months. While both of these types of feedback reports were distributed during the demonstration, a comparison between concurrent order-level feedback and aggregate retrospective feedback reports was not a primary objective of the demonstration or the evaluation.

Nevertheless, the MID evaluation has identified some important lessons about DSS feedback.

12.2.A. Value of Real-Time "Live" Feedback

The optimal implementation of the real time DSS appropriateness feedback is concurrent with the order (meaning it is delivered within seconds of placing the order). While this concurrent feedback of appropriateness rating was a requirement of MID, it was not implemented in several practices of one convener. In this practice, the order entry data including the reason for order was conveyed to staff off site. These staff received the appropriateness ratings and fed it back to the ordering clinicians. However, the feedback was often received by the ordering clinician days after the initial order was placed—sometimes after the image had

been rendered to the patient (even when the rating had been inappropriate). The optimal implementation of real-time DSS feedback also involves clinicians having easy access to supportive evidence associated with the guideline's level of evidence and level of agreement simultaneously with receipt of appropriateness feedback.

Many clinicians and conveners expressed the belief that optimal real-time, order-specific feedback would not interrupt the ordering provider with feedback if the order was appropriate. They believe the clinician's workflow should not be interrupted to confirm appropriate ordering. Other clinicians express appreciation for occasionally receiving positive feedback as long as the feedback of the appropriate rating does not slow their workflow. The MID evaluation (EMID) did not systematically test whether clinicians value real-time appropriateness feedback for inappropriate orders more than they value it for equivocally rated orders. Conveners and clinicians vary in their views on this topic. EMID also did not systematically evaluate the feasibility or desirability of providing additional prompts in the form of real time feedback when specified reasons for orders did not link with a rated guideline.

Within MID, DSS real-time feedback was not delivered according to these ideal characteristics. Clinicians and staff were aware of the limitations of the feedback and spoke freely about these limitations during focus groups. Additionally, conveners repeatedly voiced frustration with the DSS protocols that were implemented with MID. Many clinicians were not receptive to DSS feedback.

As reported in Chapter 6, clinicians often had difficulty realizing the value of the DSS as it was presented. They noted the feedback was not detailed enough and came too late. They did not find the guidelines useful in decisionmaking; a problem for many was that the DSS provided feedback after the order was placed, rather than before (Section 6.2). Clinicians were also concerned about the comprehensiveness, clarity, and validity of the guidelines (Section 6.3) and many reported not seeing or accessing any guidelines (Section 6.4).

A disadvantage of real-time feedback (relative to aggregate feedback reports) is that it focuses on a particular incident/advanced imaging order rather than providing a big-picture view of the clinician's ordering practices. If the clinician rejects the immediate feedback (because it isn't useful, doesn't seem to apply, is vague or unclear, etc.), then no learning takes place—other than, perhaps, learning how to "game" the system better next time. In this sense, an opportunity exists to influence decisions in real time in close enough proximity to the order that it could be canceled or changed. There is also the possibility of a lot of wasted feedback that clinicians don't use.

Section 2.5 showed that the live feedback facilitates decisionmaking when clear and digestible guidelines are provided and when alternative procedures to the original advanced image order are also provided. Feedback seemed particularly useful in leading to changes in procedure types (vs. changes in contrast) (Section 2.5). Only two conveners had even modest rates of cancellation in response to immediate feedback (Section 2.5); however, patterns of cancellation rates across conveners (Fig. 2.8) suggest that, for at least a subset of clinicians affiliated with two conveners, DSS feedback produced behaviors that were consistent with the intent of the demonstration.

Clinicians did express interest in receiving information that they could use in deciding whether to order an advanced image—and, if so, which specific order to place. However, they

were often frustrated by the receipt of an equivocal or inappropriate feedback message after an order was placed. While clinicians had the opportunity with MID to cancel or change orders after they had been placed and after they received the appropriateness ratings, the process was so time-consuming, especially for the many providers whose EHRs were not linked with their DSSs, that many clinicians reported ignoring the appropriateness feedback by not making a change (even if they recognized value in doing so). Once these software and DSS design–related issues are addressed so that response does not delay the clinician's workflow, real-time feedback reports showing appropriateness ratings may provide more value to clinicians. This will be even more likely if supportive evidence for the appropriateness rating is readily available to clinicians even in the context of their busy clinical practices. In fact, Chapter 6 documents that many clinicians did not even know how to find evidence associated with the appropriateness ratings. Those who could access the information were disappointed that the feedback was designed to be more encyclopedic than a desired concise summary of information that could help the clinician with decisionmaking or help the clinician explain reasons for ordering or not ordering to patients.

As documented in Chapter 6, clinicians were most enthusiastic about real-time feedback that would help them as they were deciding whether to order an advanced image. They contrasted this enthusiasm with their frustration receiving feedback after orders were placed.

Since both the aggregate and the real-time, order-specific feedback reports are retrospective, they both have limitations. Nevertheless, they each have value for clinicians.

12.2.B. Implementation and Utility Associated with Retrospective Feedback Reports

Both the real-time and the aggregate retrospective feedback reports contribute to the overall body of knowledge the clinician possesses when making a decision. Additionally, the broader trends and collected data that might be presented in such a report could be extremely useful for helping to change and improve clinician ordering over the long term—where the goal is to change patterns and habits, rather than just focusing on an individual order (which is what happens with the immediate feedback). Within MID, the basic format of aggregate feedback reports was prepared using a format agreed upon by the Lewin Group, RAND, and conveners. However, conveners were encouraged to enhance the format so that it would be most helpful to their practices. In fact, enhancements were rarely used. The day-to-day MID operations for which conveners were responsible was demanding enough that conveners did not systematically enhance the feedback reports.

Lewin has documented the distribution of feedback reports showing that most practices and providers received no practice- or person-level reports beyond the one distributed during the baseline-intervention period. While conveners initially planned to distribute aggregate reports widely and use them as tools for improvement, these plans did not come to fruition since the large number of orders that were not associated with guidelines meant that most practices and practitioners did not meet the specified CMS criteria for the volume of rated images required to receive an aggregate report. The low volume of orders placed by most clinicians made it technically infeasible to expose clinicians to enough of their ordering history to be meaningful. As noted in Table 2.3, only 10.2 percent of generalist physicians, 15.7 percent of surgical

physicians, and 21.4 percent of medical specialist physicians ordered more than 50 MID images on average during the 24-month demonstration.

While some conveners conducted meetings to help clinicians, staff, and administrators understand how to make use of the aggregate feedback reports, the reports appear not to have been an important tool for education, reflection, or engagement as they has been expected to be.

Prospective decision support provided at the time a decision is about to be made regarding a procedure can be supplemented by a suite of additional types of interventions. When clinicians have not provided the software with adequate data to allow the algorithm to link to an existing guideline, the software can request additional data from the clinicians. When the software links the reason for the order with more than one order, the software can prompt additional data queries to clarify the most relevant reason for the order at this time. When the reason for the order is not well formulated by the clinician who believes an advanced image could be helpful even though s(he) is unable to specify a reason for order that maps to an existing guideline, the software can prompt the clinician to obtain other nonradiologic or radiologic data that can guide decisionmaking. When the clinician decides an advanced image order should be placed, the software can guide the clinician to the selection of the image most likely to improve the patient's outcomes.

12.3. Recommendations

- Develop strategies to assure that all real-time feedback is delivered immediately after the order is placed.
- Initiate efforts to consider systems for prospective decision support to help clinicians decide whether to place an order before actually doing so.
- Develop guidelines to be more comprehensive and clear. Guideline validity and level of agreement should be readily available as a supplement to feedback reports when clinicians want to review them.
- Alternative radiologic and nonradiologic interventions should be readily available when feedback reports indicate inappropriate or equivocal ratings.
- Decision support help to refine the reason for order should be readily available when feedback reports indicate the order is not rated.
- Future DSS implementation efforts should realistically consider the nature of existing guidelines to estimate the proportion of orders likely to be linked to guidelines.
- Sample-size calculations should be performed after consideration of the number of orders expected to be rated and after consideration of variation in ordering volume by specialty type and image type.
- Systematic programs should be implemented to use both aggregate and real-time, order-specific retrospective feedback reports as tools for better understanding the evidence for advanced imaging use improving or harming patient outcomes.

13. Recommendations About Strategies for Motivating Physicians to Comply with Ordering Advanced Imaging Appropriately According to Appropriateness Criteria

This chapter presents recommendations to address the statute question:

- In what ways can physicians be motivated—including financial incentives—to comply with ordering advanced imaging appropriately according to appropriateness criteria?

13.1. Summary Response to Statute Question

MID was structured such that physicians would maintain their usual reimbursement rates for ordering images during the demonstration; financial incentives were not provided for ordering more or fewer images or more or less appropriate images. Even without financial incentives, most clinicians were conceptually interested in learning how to improve ordering patterns. As described in Section 6.2, this suggests a willingness by ordering clinicians to participate that might be enhanced further with some sort of incentive.

13.2. Evidence

13.2.A. Clinicians Prefer Feedback While Making Decisions in Advance of Potential Ordering Compared with Feedback After Ordering

A key issue for physicians (Section 6.2) appears to be getting feedback from the DSS before the order is placed, rather than after. Many clinicians express a preference for receiving feedback while they are still in the decisionmaking mode. This seems very important in terms of clinician motivation and engagement. Physicians would be more readily motivated to use a system that is going to help them do what they need to do (select the appropriate order or other procedure). In contrast they seem less likely to be motivated by feedback that is (a) critical of a decision they have already made (thus challenging their existing knowledge) and (b) difficult to use (requiring too much time to read and process while not systematically offering clear alternatives).

13.2.B. Limitations in Implementing MID

As noted in Chapters 8–13, the implementation of MID was limited in terms of (1) comprehensiveness and quality of the guidelines, (2) ease of use and comprehensiveness of strategies for identifying the reasons for orders, (3) metrics for reporting appropriateness that motivated change among ordering clinicians and staff, and (4) physician engagement. As noted

in Chapters 8–12, specific strategies can be implemented in response to these findings. These are likely to serve as motivational tools for physicians to further their engagement with DSSs.

13.2.C. Potential Opportunities for Improving Engagement of Clinicians and Patients

An additional, potentially motivating element for physicians is found in the patient focus group discussions presented in Chapter 7. Among focus group participants, patients seem to have a generally favorable impression of doctors' use of computers—particularly when the computer is the source of information that is useful for the patient's own care. It is conceivable that CMS could design incentives that further engage patients using a DSS to improve communication between the doctor and the patient, and to better inform both patients and their physicians.

13.2.D. CMS Should Seek Options that Save Time, Rather than Require More Time

Since time is such a valuable commodity for clinicians, renewed efforts should be implemented to unload the time burdens that clinicians have experienced as DSS, order entry, and EHRs have substantially increased administrative loads. Involving clinicians in data gathering and problem solving about time burdens is likely to be productive; their perspective on balancing workflow and patient care can reveal insights not apparent to others.

13.3. Recommendations

- Improve the comprehensiveness and quality of the guidelines.
- Enhance strategies for identifying the reasons for orders.
- Involve practicing clinicians in decisions about how to frame appropriateness ratings so they seem more like decision aids and less like report cards.
- Pay attention to strategies that can save physicians time in their workflow; think hard before burdening clinicians with responsibilities that slow down their clinical activities.
- Pilot test innovative models designed to improve clinical care involving clinicians and other stakeholders who will use this system in the planning.
- Develop strategies to align patients with physicians in enhancing the appropriate use of advanced imaging.

Section VI: Conclusion

This section is comprised of Chapter 14. It summarizes the findings of this evaluation and presents a set of recommendations.

This page is intentionally blank.

14. Conclusions

The sizable growth in utilization of advanced imaging during the last decade for Medicare FFS beneficiaries prompted CMS to examine reasons for the growth and strategies for controlling the growth of services that were not clinically appropriate. The MID instructs the U.S. Department of Health and Human Services to collect data on Medicare FFS patients using DSSs to determine the appropriateness of services in relation to established criteria and to provide participating physicians with feedback reports that permit comparison against physician peers on adherence to appropriateness criteria.

This report summarizes data associated with analyses of claims, decision support data, and focus groups. From claims data, we see evidence for a drop in utilization that is significant from the predemonstration to the demonstration period for all but one of the seven conveners. However, this drop does not differ significantly for demonstration versus comparison sites for five of seven conveners. Of seven conveners studied, two showed statistically significant reductions in the number of advanced images utilized by the demonstration compared with the matched control providers. The estimated differences were relatively small, implying an additional reduction of one MID image per 100 Medicare beneficiaries within demonstration versus control sites. Given that even these two significant results are sensitive to whether images are included in months when no beneficiaries are found for a particular provider, we consider the evidence for these demonstration effects to be weaker than the confidence intervals imply on their own. In whole, our claims analysis provides no evidence that appropriateness feedback leads to anything beyond a small reduction—if any reduction at all—in MID image volume.

Examination of the DSS data showed that almost two-thirds of MID advanced image orders placed through a DSS were not linked to a guideline, meaning the ordering clinicians or staff did not receive appropriateness feedback for the majority of orders placed. This pattern had a major impact on diluting any potential effect of the demonstration. Among rated orders, we observed an increase in the proportion of orders that were rated as appropriate during the intervention compared with the baseline period. Analyses combining changes in the frequency of rated orders and changes in appropriateness show that appropriateness levels did improve substantially for rated orders after DSS feedback was turned on. However, a very large proportion of orders were not rated and the fraction of unrated orders increased for two conveners (across all specialties) and within particular specialties for all but two of the conveners. Although these results suggest that providers do adjust advanced imaging ordering practices when given DSS feedback—or at least, adjust their explanations of the reasons for the order—it is not clear how these changes would affect total appropriateness rates of all ordered images.

Detailed analyses from DSS data provide evidence that clinicians do respond to aspects of the decision support intervention, particularly to real-time feedback regarding alternative

procedures. This provides an empirical model showing the DSS can motivate change through the expected mechanism. Furthermore, analyses of trends in the appropriateness of orders for advanced imaging, which was limited to orders that received appropriateness ratings, showed improvements over time of about 7 percentage points. Six of seven conveners achieved higher rates of appropriate ordering, while the seventh convener already had the highest rate of appropriate orders at baseline (82 percent). These increases in appropriateness are indicative of a successful intervention to the extent that the proportion of orders that are successfully rated by DSS remains stable between periods.

Despite the inclusion of orders from more than 5,000 clinicians, nearly two-thirds of clinicians in our sample placed fewer than 20 orders—suggesting that many clinicians might not have been adequately exposed to the intervention to influence their ordering behavior.

14.1. Summary of Evaluation Findings Associated with 11 MID Research Questions

Chapters 2–13 of this report present analysis of RAND's findings associated with 11 questions that emerged from the congressional statute pertinent to the MID. Table 14.1 summarizes RAND's findings related to these questions.

Table 14.1. Summary of Findings to 11 Evaluation Questions

Question 1: What were the rates of appropriate, uncertain, and inappropriate advanced imaging orders over the course of the demonstration? (Chapter 2)
• A majority (nearly 2/3) of advanced imaging orders in both the baseline and intervention periods were not assigned an appropriateness rating by the DSSs because they could not be linked to clinical guidelines. • Among orders receiving an appropriateness rating, 79.5 percent were rated appropriate across all conveners over the baseline and intervention period, while only 13.1 percent were rated equivocal and 7.4 percent were rated inappropriate.
Question 2: Were any patterns or trends evident in the appropriateness or inappropriateness of advanced imaging procedure orders? (Chapter 3)
• The proportion of rated orders that were found appropriate increased in the intervention period compared to the baseline period. However, a very large proportion of orders were not rated.
Question 3: Is there a relationship between the appropriateness of advanced imaging procedure orders and imaging results? (Chapter 4)
• Without the opportunity to access data describing imaging test results, we elected to examine whether receipt of patterns of appropriateness feedback impacts the subsequent volume of advanced image orders. The providers whose percent Inappropriate orders improved during early intervention substantially outnumbered those whose percent worsened. This result was sustained.
Question 4: Were any national or regional patterns or trends evident in utilization of advanced imaging procedures? (Chapter 5)
• There was no evidence that the intervention had a meaningful effect on advanced imaging volume.

Table 14.1—Cont.

Question 5: How satisfied were physicians in the demonstration with being exposed to advanced imaging appropriateness criteria? (Chapter 6)
• Clinician focus group participants reported that they were variably satisfied with the demonstration project, with some describing their support for the "idea" behind the demonstration. However, most believed that more pilot testing should have been implemented prior to the implementation of a national demonstration. • Most did not believe that the available guidelines or the software linking the user interface and guidelines were well enough developed at this time to adequately engage clinicians in a meaningful way.
Question 6: How satisfied were Medicare patients in the demonstration with receiving an advanced imaging procedure after a physician was exposed to appropriateness criteria? (Chapter 7)
• According to physician focus group participants, the DSS implementation did not have any substantial impact on clinicians' relationships with patients or patient satisfaction.
Question 7: Was the system for determining appropriateness in the demonstration acceptable for identifying appropriate advanced imaging orders versus inappropriate ones? (Chapter 8)
• The use of DSSs for identifying appropriate versus inappropriate advanced imaging orders in the demonstration was not generally acceptable to clinicians in the form that was deployed. • Furthermore, the system was not efficient, as it did not assign an appropriateness rating to 2/3 of the ordered advanced images.
Question 8: Would exposing physicians to advanced imaging appropriateness criteria at the time of order affect the volume of utilization? (Chapter 9)
• In the demonstration, exposing physicians to advanced imaging appropriateness criteria at the time of order did not substantially impact volume.
Question 9: If expanded to a broader population of Medicare beneficiaries, should physicians who demonstrate that their ordering patterns are consistently appropriate be exempt from requirements to consult appropriateness criteria? (Chapter 11)
• A decision theoretic approach provides a model of what types of ordering histories should exempt providers from DSS use and for how long. However, the meaningful application of the model is predicated by the assumption that DSSs are broadly useful in the setting where the exemption criteria would be considered.
Question 10: To what extent is live feedback on the appropriateness of advanced imaging orders from a decision support system better or worse than feedback reports to individual physicians or physician practices? (Chapter 12)
• In the demonstration, exposing physicians to advanced imaging appropriateness criteria at the time of order did not substantially affect volume.
Question 11: In what ways can physicians be motivated—including financial incentives—to comply with ordering advanced imaging appropriately according to appropriateness criteria? (Chapter 13)
• Even without financial incentives, most clinicians were conceptually interested in learning how to improve ordering patterns suggesting a willingness by ordering clinicians to participate that might be enhanced further with some sort of incentive.

14.2. Recommendations

The remainder of this chapter presents leading recommendations for enhancing the effectiveness of DSSs for improving the appropriateness of advanced imaging. More detailed versions of these recommendations are found in Chapters 8–13.

We need a more comprehensive set of guidelines that can cover a substantially greater proportion of advanced images ordered for Medicare beneficiaries.

- These guidelines need to more explicitly map the clinical characteristics that bedside clinicians are familiar with to the distinguishing features of the clinical guidelines.
- Alternative clinical and procedural options should be made more explicit and should include both radiologic and nonradiologic options.
- The guidelines need to consistently and explicitly document level of evidence, as is standard for most published guidelines.
- Guideline development should not be restricted to those from the specialty society associated with the procedure but should include other respected bodies less likely to be subject to potential or assumed bias.
- When guidelines are not available, more information should be shared about the reasons why.
- Additional evidence is needed about the best methods for incorporating level of certainty into the guidelines and for making the rating system reflect that evidence.
- Guidelines need to be more explicit about how appropriateness might vary for initial compared with serial images. The appropriateness of serial images when signs and symptoms are escalating or resolving need to be distinguished from the timing of serial images when patients are stable. Guidelines need to be more explicit about periodicity.
- Strategies should be developed to align patients with physicians in enhancing the appropriate use of advanced imaging.

Methods for communicating feedback to clinicians need further exploration.

- More research is needed on DSS design to better understand circumstances under which DSS feedback is most effective. In particular, it is important to understand advantages and disadvantages of providing feedback based on decision support prior to decisions being formulated and implemented, as compared with afterward.
- While an ordinal scale for measuring appropriateness has been used in many settings, this strategy should be further evaluated, especially when evidence is limited.

DSS should help clinicians perform their work more efficiently and effectively.

- DSS should facilitate easy order entry, specification of the reason for order, receipt of appropriate orders, and awareness of alternative interventions (radiologic or nonradiologic) that could improve patients presenting with the problem the clinician is assessing.
- More effort is needed to design easy-to-access guidelines, ratings, and alternatives.
- Further software linkages between clinical data and EHR clinical details need to be implemented.

Expertise within and across multiple disciplines will be required to address the current challenges to effective DSS use.

- While teams with discipline-specific expertise should explore options for addressing challenges, cross-disciplinary teams will also be needed to align priorities essential for aligning clinical, information technology, administration, and patient stakeholders.
- Clinical inputs must be responsive to the workflow requirements of the set of generalist and specialist physicians who traditionally have been responsible for advanced image ordering, as well as to the growing set of ancillary and support staff who implement many tasks previously performed only by clinicians.
- DSS structures and processes should be further integrated into training curricula for clinicians using DSSs.
- Informaticists should align with those developing DSSs for advanced imaging.
- Radiologists and clinicians should consider how their workflows are pertinent to order entry and documenting reasons for orders can be aligned.
- Protocols for assigning reasons for ordering advanced images should be known to those involved with ordering. These protocols should be consistent with both national and local guidelines.
- Radiology protocolling procedures should be aligned with DSS protocols used by clinicians and staff to enhance workflow, coordination of taxonomies for specifying reasons for orders, and special needs associated with ordered images or associated patients.
- Discrepancies noted between national and local guidelines, as well as radiology and other specialty guidelines, should be explored. Reasons should be specified so that relevant stakeholders understand reasons for ongoing discrepancies.
- Clinical leaders should engage with guidelines and model their use for clinicians and staff. Areas of guideline concern or disagreement should be discussed so strategies for improving DSSs can be developed in a manner satisfactory to users.
- When DSSs are used, they should be actively managed such that clinicians associated with orders that are not linked to guidelines or that frequently place inappropriate orders should engage in discussions about the reasons for this. These reasons should be vetted with other clinicians, radiologists, and administration. Administration should consider discussing these issues with payers to further discriminate valid versus other reasons for discrepancies.

Specific efforts to support clinicians as they undergo major changes in health care delivery should be provided.

- Providing adequate training for clinicians, attending to how workflow could be affected, and supporting clinicians as their productivity is challenged can make an enormous difference in the degree to which clinicians buy in to the process.
- Problems should be identified and addressed using a multidisciplinary team as needed to assure that the clinical, administrative, patient engagement, fiscal, and information technology issues are aligned. In particular, a trend is emerging for clinicians to work in teams with others so that the team is effective, while each individual works at the "top of their license." Attention should be paid to how DSSs could function when the tasks of deciding "what to order" and "ordering" are configured as distinct tasks to be carried out by different individuals in some practice settings.

Identify the specific DSS challenges associated with subpopulations

- Given the complexity of addressing the many challenges to DSS that have been identified, it would be helpful to pinpoint specific subpopulations that could benefit from specific interventions.
- Stratifying scale-up efforts to focus on guidelines that are well specified, well supported with evidence, and consistently agreed upon would facilitate an evaluation that could be more nuanced in determining which aspects of the DSS intervention make a difference.
- Working with guideline developers to ensure that guideline ratings are available for both first-time and follow-up orders, with each set being supported with evidence, will provide clinicians a better understanding of when imaging can improve their patient's outcomes.
- Given the many factors that have been identified as influencing the clinician's and practices' experiences with DSSs, scaling up with a more focused approach could allow the process to reach an earlier tipping point of acceptability.

14.3. Conclusion

In summary, we found no evidence that the intervention led to anything beyond a small reduction—if any reduction at all—in advanced imaging volume. Furthermore, since more than half of the ordered advanced images did not link with guidelines, in these instances, appropriateness feedback was not available to share with ordering clinicians. Despite these limitations, detailed analyses from DSS data reveal that clinicians do respond to concurrent feedback showing them imaging strategies that can serve as alternatives to their original orders. The appropriateness of advanced imaging orders that were assigned appropriateness ratings increased by about 7 percentage points between the baseline and intervention periods. These increases in appropriateness are indicative of a successful intervention to the extent that advanced imaging orders can be successfully rated for appropriateness.

While the MID evaluation provides support for changes in appropriateness in association with clinician exposure to guidelines, the conceptual model for changing behavior should be expanded to further attend to the needs of clinicians who serve as the vector for change. At the same time, clinicians and patients should be better integrated into protocols for developing guidelines, appropriateness ratings, assessment of agreement associated with ratings, and implementation of guidelines into clinical workflows. These efforts will need to be supplemented with enhanced information technology advances that will leverage the clinical detail that is increasingly populating EHRs and administrative data sets.

The observed increases in appropriateness are small but encouraging, especially in light of the many challenges experienced with MID's implementation. We expect that, with the implementation of our recommendations, many of the identified obstacles can be overcome.

This page is intentionally blank.

Bibliography

ABIM Foundation. (undated-a). *About Choosing Wisely.* Retrieved from http://choosingwisely.org/?page_id=8

ABIM Foundation. (undated-b). Choosing Wisely home page. Retrieved from http://www.choosingwisely.org

ACR—*See* American College of Radiology.

American College of Radiology. (2010). *About Us.* Retrieved January 29, 2014 http://www.imagewisely.org/About-Us

———. (2014). *National Radiology Data Registry.* Retrieved July 15, 2014: http://www.acr.org/Quality-Safety/National-Radiology-Data-Registry

Bautista, A. B., Burgos, A., Nickel, B. J., Yoon, J. J., Tilara, A. A., Amorosa, J. K., & American College of Radiology, Appropriateness. (2009). Do clinicians use the American College of Radiology appropriateness criteria in the management of their patients? *AJR Am J Roentgenol, 192*(6), 1581–1585. doi: 10.2214/AJR.08.1622

Begun, J. W., Riley, W. J., & Hodges, J. S. (2013). Exploratory Analysis of High CT Scan Utilization in Claims Data. *J Am Coll Radiol*. doi: 10.1016/j.jacr.2013.04.011

Bernardy, M., Ullrich, C. G., Rawson, J. V., Allen, B., Jr., Thrall, J. H., Keysor, K. J., . . . Mabry, M. R. (2009). Strategies for managing imaging utilization. *J Am Coll Radiol, 6*(12), 844–850. doi: 10.1016/j.jacr.2009.08.003

Blachar, A., Tal, S., Mandel, A., Novikov, I., Polliack, G., Sosna, J., . . . Shemer, J. (2006). Preauthorization of CT and MRI examinations: Assessment of a managed care preauthorization program based on the ACR appropriateness criteria and the Royal College of Radiology guidelines. *J Am Coll Radiol, 3*(11), 851–859. doi: 10.1016/j.jacr.2006.04.005

Blackmore, C. C., Mecklenburg, R. S., & Kaplan, G. S. (2011). Effectiveness of clinical decision support in controlling inappropriate imaging. *J Am Coll Radiol, 8*(1), 19–25. doi: 10.1016/j.jacr.2010.07.009

Blackmore, C. C., & Medina, L. S. (2006). Evidence-based radiology and the ACR Appropriateness Criteria. *J Am Coll Radiol, 3*(7), 505-509. doi: 10.1016/j.jacr.2006.03.003

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and quantity, 36*, 391–409.

Bowen, S., Johnson, K., Reed, M. H., Zhang, L., & Curry, L. (2011). The effect of incorporating guidelines into a computerized order entry system for diagnostic imaging. *J Am Coll Radiol, 8*(4), 251–258. doi: 10.1016/j.jacr.2010.11.020

Brenner, D. J., & Hall, E. J. (2007). Computed tomography—an increasing source of radiation exposure. *N Engl J Med, 357*(22), 2277–2284. doi: 10.1056/NEJMra072149

Brink, J. A., & Amis, E. S., Jr. (2010). Image Wisely: A campaign to increase awareness about adult radiation protection. *Radiology, 257*(3), 601–602. doi: 10.1148/radiol.10101335

Brook, R. H., Chassin, M. R., Fink, A., Solomon, D. H., Kosecoff, J., & Park, R. E. (1986). A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care, 2*(1), 53-63.

Brook, R. H., Chassin, M. R., Fink, A., Solomon, D. H., Kosecoff, J., & Park, R. E. (1991). *A Method for the Detailed Assessment of the Appropriateness of Medical Technologies.* Santa Monica, CA: RAND Corporation, N-3376-HHS. As of July 16, 2014: http://www.rand.org/pubs/notes/N3376.html

Burke, J. F., Kerber, K. A., Iwashyna, T. J., & Morgenstern, L. B. (2012). Wide variation and rising utilization of stroke magnetic resonance imaging: data from 11 states. *Ann Neurol, 71*(2), 179-185. doi: 10.1002/ana.22698

Chassin, M. R., Brook, R. H., Park, R. E., Keesey, J., Fink, A., Kosecoff, J., . . . Solomon, D. H. (1986). Variations in the use of medical and surgical services by the Medicare population. *N Engl J Med, 314*(5), 285-290. doi: 10.1056/NEJM198601303140505

Curry, L., & Reed, M. H. (2011). Electronic decision support for diagnostic imaging in a primary care setting. *J Am Med Inform Assoc, 18*(3), 267–270. doi: 10.1136/amiajnl-2011-000049

Deyo, R. A. (2002). Cascade effects of medical technology. *Annu Rev Public Health, 23*, 23–44. doi: 10.1146/annurev.publhealth.23.092101.134534092101.134534

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *Am J Community Psychol, 41*, 327–350.

Fazel, R., Krumholz, H. M., Wang, Y., Ross, J. S., Chen, J., Ting, H. H., . . . Nallamothu, B. K. (2009). Exposure to low-dose ionizing radiation from medical imaging procedures. *N Engl J Med, 361*(9), 849–857. doi: 10.1056/NEJMoa0901249

Fereday, J., & Muir-Cochrane, E. (2008). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods, 5*, 80–92.

Field, M. J., & Lohr, K. N. (1990). *Clinical practice guidelines: Directions for a new program*. Washington, D.C.: National Academy Press.

Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. (1984). Consensus methods: Characteristics and guidelines for use. *Am J Public Health, 74*(9), 979–983.

Fisher, E. S. (2003). Medical care—Is more always better? *N Engl J Med, 349*(17), 1665–1667. doi: 10.1056/NEJMe038149

Fitch, K., Bernstein, S., Aguilar, M., & Burnand, B. (2001). *The RAND/UCLA appropriateness method user's manual 2001*. Santa Monica, CA: RAND Corp.

Georgiou, A., Prgomet, M., Markewycz, A., Adams, E., & Westbrook, J. I. (2011). The impact of computerized provider order entry systems on medical-imaging services: A systematic review. *J Am Med Inform Assoc, 18*(3), 335–340. doi: amiajnl-2010-000043 [pii]

Gibbons, R. J., Miller, T. D., Hodge, D., Urban, L., Araoz, P. A., Pellikka, P., & McCully, R. B. (2008). Application of appropriateness criteria to stress single-photon emission computed tomography sestamibi studies and stress echocardiograms in an academic medical center. *J Am Coll Cardiol, 51*(13), 1283–1289. doi: 10.1016/j.jacc.2007.10.064

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Soc Probl, 12*, 436–445.

Gliwa, C., & Pearson, S. D. (2014). Evidentiary rationales for the Choosing Wisely Top 5 lists. *JAMA, 311*(14), 1443–1444. doi: 10.1001/jama.2013.285362

Grilli, R., Magrini, N., Penna, A., Mura, G., & Liberati, A. (2000). Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet, 355*(9198), 103–106. doi: 10.1016/S0140-6736(99)02171-6

Grol, R., & Grimshaw, J. (2003). From best evidence to best practice: Effective implementation of change in patients' care. *Lancet, 362*(9391), 1225–1230. doi: 10.1016/S0140-6736(03)14546-1

Harris, Gardiner. (March 26, 2010). More doctors giving up private practices, *New York Times*. Retrieved from http://www.nytimes.com/2010/03/26/health/policy/26docs.html?_r=2

Hendel, R. C., Berman, D. S., Di Carli, M. F., Heidenreich, P. A., Henkin, R. E., Pellikka, P. A., et al. Society of Nuclear, Medicine. (2009). ACCF/ASNC/ACR/AHA/ASE/SCCT/SCMR/SNM 2009 Appropriate Use Criteria for Cardiac Radionuclide Imaging: A Report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, the American Society of Nuclear Cardiology, the American College of Radiology, the American Heart Association, the American Society of Echocardiography, the Society of Cardiovascular Computed Tomography, the Society for

Cardiovascular Magnetic Resonance, and the Society of Nuclear Medicine. *J Am Coll Cardiol, 53*(23), 2201-2229. doi: 10.1016/j.jacc.2009.02.013

Hendel, R. C., Cerqueira, M., Douglas, P. S., Caruth, K. C., Allen, J. M., Jensen, N. C., . . . Wolk, M. (2010). A multicenter assessment of the use of single-photon emission computed tomography myocardial perfusion imaging with appropriateness criteria. *J Am Coll Cardiol, 55*(2), 156–162. doi: 10.1016/j.jacc.2009.11.004

Higashi, T., Shekelle, P. G., Adams, J. L., Kamberg, C. J., Roth, C. P., Solomon, D. H., Reuben, D. B., Chang, L., MacLean, C. H., Chang, J. T., Young, R. T., Saliba, D. M., Wenger, N. S. (2005). Quality of care is associated with survival in vulnerable older patients. *Annals of Internal Medicine 143*, 274–281.

Hillman, B. J., & Goldsmith, J. (2010a). Imaging: the self-referral boom and the ongoing search for effective policies to contain it. *Health Aff (Millwood), 29*(12), 2231-2236. doi: 10.1377/hlthaff.2010.1019

Hillman, B. J., & Goldsmith, J. C. (2010b). The uncritical use of high-tech medical imaging. *N Engl J Med, 363*(1), 4-6. doi: 10.1056/NEJMp1003173

Institute of Medicine, Committee to Advise the Public Health Service on Clinical Practice Guidelines, (Ed.). (1990). *Clinical practice guidelines: Directions of a new program*. Washington, D.C.: National Academy Press.

Ip, I. K., Schneider, L. I., Hanson, R., Marchello, D., Hultman, P., Viera, M., . . . Khorasani, R. (2012). Adoption and meaningful use of computerized physician order entry with an integrated clinical decision support system for radiology: Ten-year analysis in an urban teaching hospital. *J Am Coll Radiol, 9*(2), 129–136. doi: 10.1016/j.jacr.2011.10.010

Jamal, T., & Gunderman, R. B. (2008). The American College of Radiology appropriateness criteria: The users' perspective. *J Am Coll Radiol, 5*(3), 158–160. doi: 10.1016/j.jacr.2007.08.021

Jamtvedt, G., Young, J. M., Kristoffersen, D. T., Thomson O'Brien, M. A., & Oxman, A. D. (2003). Audit and feedback: Effects on professional practice and health care outcomes. *Cochrane Database Syst Rev*(3), CD000259. doi: 10.1002/14651858.CD000259

Kahn, K. L., Khodyakov, D., Weidmer, B., Wenger, N. S., Brantley, I., Burgette, L., . . . Hussey, P. S. (2013). *Medicare Imaging Demonstration Evaluation: Clinician and Staff Focus Group Report*. Submitted to CMS.

Kahn, K. L., Kosecoff, J., Chassin, M. R., Flynn, M. F., Fink, A., Pattaphongse, N., . . . Brook, R. H. (1988a). Measuring the clinical appropriateness of the use of a procedure. Can we do it? *Med Care, 26*(4), 415–422.

Kahn, K. L., Kosecoff, J., Chassin, M. R., Solomon, D. H., & Brook, R. H. (1988b). The use and misuse of upper gastrointestinal endoscopy. *Ann Intern Med, 109*(8), 664–670.

Kahn, K. L., Rogers, W. H., Rubenstein, L. V., Sherwood, M. J., Reinisch, E. J., Keeler, E. B., . . . Brook, R. H. (1990). Measuring quality of care with explicit process criteria before and after implementation of the DRG-based prospective payment system. *JAMA, 264*(15), 1969–1973.

Kahn, K. L., Tisnado, D. M., Adams, J. L., Liu, H., Chen, W. P., Hu, F. A., Damberg, C. L. (2007). Does ambulatory process of care predict health-related quality of life outcomes for patients with chronic disease? *Health Serv Res, 42*(1 Pt 1), 63–83. doi: 10.1111/j.1475-6773.2006.00604.x

Kennedy, S., & Forman, H. P. (2012). Deficit reduction act: Effects on utilization of noninvasive musculoskeletal imaging. *Radiology, 264*(1), 146–153. doi: 10.1148/radiol.12110993

Lang, K., Huang, H., Lee, D. W., Federico, V., & Menzin, J. (2013). National trends in advanced outpatient diagnostic imaging utilization: an analysis of the medical expenditure panel survey, 2000–2009. *BMC Med Imaging, 13*, 40. doi: 10.1186/1471-2342-13-40

Lee, C. I., Haims, A. H., Monico, E. P., Brink, J. A., & Forman, H. P. (2004). Diagnostic CT scans: Assessment of patient, physician, and radiologist awareness of radiation dose and possible risks. *Radiology, 231*(2), 393–398. doi: 10.1148/radiol.2312030767

Lee, D. W., Duszak, R., Jr., & Hughes, D. R. (2013). Comparative analysis of Medicare spending for medical imaging: Sustained dramatic slowdown compared with other services. *AJR Am J Roentgenol, 201*(6), 1277–1282. doi: 10.2214/AJR.13.10999

Lee, D. W., & Levy, F. (2012). The sharp slowdown in growth of medical imaging: An early analysis suggests combination of policies was the cause. *Health Aff (Millwood), 31*(8), 1876–1884. doi: 10.1377/hlthaff.2011.1034

Lehnert, B. E., & Bree, R. L. (2010). Analysis of appropriateness of outpatient CT and MRI referred from primary care clinics at an academic medical center: How critical is the need for improved decision support? *J Am Coll Radiol, 7*(3), 192–197. doi: 10.1016/j.jacr.2009.11.010

Levin, D. C., Rao, V. M., & Parker, L. (2010). Physician orders contribute to high-tech imaging slowdown. *Health Aff (Millwood), 29*(1), 189-195. doi: 29/1/189 [pii]

Levin, D. C., Rao, V. M., & Parker, L. (2012). The recent downturn in utilization of CT: the start of a new trend? *J Am Coll Radiol, 9*(11), 795-798. doi: 10.1016/j.jacr.2012.05.023

Levy, G., Blachar, A., Goldstein, L., Paz, I., Olsha, S., Atar, E., . . . Bar Dayan, Y. (2006). Nonradiologist utilization of American College of Radiology appropriateness criteria in a

preauthorization center for MRI requests: Applicability and effects. *AJR Am J Roentgenol, 187*(4), 855–858. doi: 10.2214/AJR.05.1055

The Lewin Group. (2011a). *Medical Society Guidelines Relevant to Medicare Imaging Demonstration Procedures.* Prepared for CMS.

———. (2011b). *Medicare Imaging Demonstration Convener Manual version 1.2.* Prepared for CMS.

———. (2010). *Design Report: Medicare Appropriateness of Use Imaging Demonstration.* Prepared for CMS.

MacQueen, K. M., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis. *Cult Anthropol Method, 10*(2), 31–36.

Maddox, T. G. (2002). Adverse reactions to contrast material: recognition, prevention, and treatment. *Am Fam Physician, 66*(7), 1229-1234.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods, 9*(4), 403–425. doi: 10.1037/1082-989X.9.4.403

McCully, R. B., Pellikka, P. A., Hodge, D. O., Araoz, P. A., Miller, T. D., & Gibbons, R. J. (2009). Applicability of appropriateness criteria for stress imaging: Similarities and differences between stress echocardiography and single-photon emission computed tomography myocardial perfusion imaging criteria. *Circ Cardiovasc Imaging, 2*(3), 213–218. doi: 10.1161/circimaging.108.798082

MedPAC—See Medicare Payment Advisory Commission.

Medicare Payment Advisory Commission. (2011). *Report to the Congress: Medicare and the health care delivery system.* Washington, DC.

———. (2013). *A data book: Healthcare spending and the Medicare program*. Washington, Dc: MedPAC.

Mehta, R., Ward, R. P., Chandra, S., Agarwal, R., & Williams, K. A. (2008). Evaluation of the American College of Cardiology Foundation/American Society of Nuclear Cardiology appropriateness criteria for SPECT myocardial perfusion imaging. *J Nucl Cardiol, 15*(3), 337–344. doi: 10.1016/j.nuclcard.2007.10.010

Miller, J. A., Raichlin, E., Williamson, E. E., McCully, R. B., Pellikka, P. A., Hodge, D. O., . . . Araoz, P. A. (2010). Evaluation of coronary CTA appropriateness criteria in an academic medical center. *J Am Coll Radiol, 7*(2), 125–131. doi: 10.1016/j.jacr.2009.08.013

Mitchell, J. M., & Lagalia, R. R. (2009). Controlling the escalating use of advanced imaging: The role of radiology benefit management programs. *Med Care Res Rev, 66*(3), 339–351. doi: 1077558709332055 [pii]

Moskowitz, H., Sunshine, J., Grossman, D., Adams, L., & Gelinas, L. (2000). The effect of imaging guidelines on the number and quality of outpatient radiographic examinations. *AJR Am J Roentgenol, 175*(1), 9-15. doi: 10.2214/ajr.175.1.1750009

Murphy, M. K., Brady, T. J., Nasir, K., Gazelle, G. S., Bamberg, F., Truong, Q. A., . . . Blankstein, R. (2010). Appropriateness and utilization of cardiac CT: Implications for development of future criteria. *J Nucl Cardiol, 17*(5), 881–889. doi: 10.1007/s12350-010-9252-0

Owens, D. K., Lohr, K. N., Atkins, D., Treadwell, J. R., Reston, J. T., Bass, E. B., . . . Helfand, M. (2010). AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions--agency for healthcare research and quality and the effective health-care program. *J Clin Epidemiol, 63*(5), 513–523. doi: 10.1016/j.jclinepi.2009.03.009

Park, R. E., Fink, A., Brook, R. H., Chassin, M. R., Kahn, K. L., Merrick, N. J., . . . Solomon, D. H. (1986). Physician ratings of appropriate indications for six medical and surgical procedures. *Am J Public Health, 76*(7), 766–772.

Park, R. E., Fink, A., Brook, R. H., Chassin, M. R., Kahn, K. L., Merrick, N. J., . . . Solomon, D. H. (1989). Physician ratings of appropriate indications for three procedures: Theoretical indications vs indications used in practice. *Am J Public Health, 79*(4), 445–447.

Rosenthal, D. I., Weilburg, J. B., Schultz, T., Miller, J. C., Nixon, V., Dreyer, K. J., & Thrall, J. H. (2006). Radiology order entry with decision support: Initial clinical experience. *J Am Coll Radiol, 3*(10), 799–806. doi: 10.1016/j.jacr.2006.05.006

Shaneyfelt, T. M., & Centor, R. M. (2009). Reassessment of clinical practice guidelines: go gently into that good night. *JAMA, 301*(8), 868–869. doi: 10.1001/jama.2009.225

Shekelle, P. G., & Schriger, D. L. (1993). *Using the Appropriateness Method in Clinical Practice Guideline Development*. Santa Monica, CA: RAND Corporation, MR-233-AHCPR. As of July 16, 2014: http://www.rand.org/pubs/monograph_reports/MR233.html

Shekelle, P. G., Kahan, J. P., Bernstein, S. J., Leape, L. L., Kamberg, C., Park, R. E. (1998). Thereproducibility of a method to identify the overuse and underuse of medical procedures. *NEJM (338)*26, pp. 1888–1895.

Shekelle, P. G., Park, R. E., Kahan, J. P., Leape, L. L., Kamberg, C., Bernstein, S. J. (2001). Sensitivity and specificity of the RAND/UCLA appropriateness method to identify the

overuse of coronry revascularization and hysterectomy. *Journal of Clinical Epidemiology, (54)*10, pp. 1004–1010

Sistrom, C. L., Dang, P. A., Weilburg, J. B., Dreyer, K. J., Rosenthal, D. I., & Thrall, J. H. (2009). Effect of computerized order entry with integrated decision support on the growth of outpatient procedure volumes: Seven-year time series analysis. Radiology, 251(1), 147–155. doi: 10.1148/radiol.2511081174

Solberg, L. I., Wei, F., Butler, J. C., Palattao, K. J., Vinz, C. A., & Marshall, M. A. (2010). Effects of electronic decision support on high-tech diagnostic imaging orders and patients. *Am J Manag Care, 16*(2), 102–106.

stata.com. (Undated). *Random-effects ordered logistic models.* Retrieved from: http://www.stata.com/manuals13/xtxtologit.pdf

U.S. Congress. (2006). *Deficit Reduction Act of 2005*. S.1932.

U.S. Congress. (2010). Affordable Care Act, Section 3003. Improvements to the Physician Feedback Program.

U.S. Department of Health and Human Services. (Undated). National Guideline Clearinghouse, home page. Retrieved from http://www.guideline.gov/

Vartanians, V. M., Sistrom, C. L., Weilburg, J. B., Rosenthal, D. I., Thrall, J. H. (2010). Increasing the Appropriateness of Outpatient Imaging: Effects of a Barrier to Ordering Low-Yield Examinations. *Radiology, 255*(3), 842–849. Doi: 10.1148/radiol. 10091228.

Winslow, C. M., Solomon, D. H., Chassin, M.R., Kosecoff, J., Merrick, N. J., and Brook, R. H. (1988). The appropriateness of carotid endarterectomy. N Engl J Med 318, 721–727. Doi: 10.1056/NEJM198803243181201

Wolters Kluwer Health. (Undated). UpToDate home page. Retrieved from: http://www.uptodate.com/home

Technical Appendix A: DSS and Claims Methods

Methods for Analyses of DSS Data

Description of the DSS Dataset

Data from DSSs were collected by The Lewin Group and transmitted to RAND on an approximately quarterly basis to facilitate ongoing analyses. The DSS data files include information collected during ROE and appropriateness scoring, including patient and ordering clinician identifiers; diagnoses, signs, and symptoms associated with imaging procedure orders; the appropriateness rating for the imaging procedure order; documentation of the original and final orders and reason from change or cancellation of the original order, if applicable (i.e., if following observation of the appropriateness rating, the ordering clinician changes or cancels the order).

While the DSS consistently collected the same data and calculated an appropriateness rating for each ordered image during both the baseline and intervention periods, the appropriateness rating was only displayed to the ordering clinician during the intervention period. In this way, the implementation and evaluation teams had a DSS-generated record of ordered images and their appropriateness ratings throughout the entire 24-month demonstration.

DSS Record Inclusion Criteria

The DSS analyses included all orders placed through a convener's DSS during the 24-month demonstration period. Because our intent was to assess the appropriateness of clinicians' decisionmaking in response to the receipt of feedback on the appropriateness of each order, we did not restrict our analysis to orders for imaging procedures that were ultimately rendered. Inclusion criteria included:

7. All DSS records for all types of ordering clinicians (physician, nurse practitioner, physician assistant, or staff)
8. All DSS records, regardless of whether they were successfully matched to a claim (the intended goal of the implementation contractor) or canceled by the ordering clinician.

Exclusion criteria included:

1. Orders placed before October 1, 2011. These appear to represent "test cases."
2. DSS records with missing appropriateness data or records flagged as incomplete by conveners.

Assignment of Clinician Specialty

We assigned a specialty to each ordering clinician using data from one of two sources. We began by using data from the NPPES to crosswalk a clinician's National Provider Identifier (NPI) that was present on each DSS record to one or more provider taxonomy codes representing unique specialties. One of these codes in NPPES may be flagged as the primary taxonomy code. If the NPPES record included a single provider taxonomy code or a provider taxonomy code flagged as primary, then that provider taxonomy code was assigned to the ordering clinician. If the NPPES record included multiple provider taxonomy codes with none flagged as primary, then we used the most specific provider taxonomy code listed or, in case of a tie, the first provider taxonomy code listed on the record. If an ordering clinician NPI was not found in NPPES, we assigned the specialty code appearing on the greatest number of carrier claims for services performed by the clinician over the course of a two-year period. We then used each clinician's assigned provider taxonomy code to create seven mutually exclusive categories: physician generalist, physician medical specialists, physician surgical specialists, nonphysician generalist, nonphysician medical specialists, nonphysician surgical specialists, and other.

Methods for Processing Combination Orders

Lewin designed the DSS data collection system to capture orders for CT abdomen and CT pelvis as a single DSS record rather than as two separate ones. In some cases, however, clinicians submitted two separate DSS orders for these procedures. One of RAND's first data-cleaning steps was to identify orders for CT abdomen and CT pelvis placed on the same day for the same patient. We consolidated these records into a single record and assigned to the consolidated record the more favorable appropriateness score from the two original records.

Methods for Identifying Duplicate Orders

One of the implications of incomplete integration of ROE and the DSS was that clinicians who wanted to change their order (for example, after receiving DSS feedback) were required to cancel their initial order and enter a second order in their ROE system. This often resulted in the creation of two DSS records. Failure to detect and remove these "duplicate" orders would distort each convener's denominator and possibly lead to erroneous inferences regarding appropriateness if, for example, clinicians tended to change to a more appropriate imaging procedure but both the initial DSS record and the final DSS record were retained in the analysis. We defined duplicate orders as multiple orders placed on the same day for the same body part for the same patient. Body parts were defined using the crosswalk found in Table A.1.

Table A.1. Crosswalk Between Body Part and MID Imaging Procedure

Body Part	MID Imaging Procedure
Brain	CT Brain, MRI Brain
Abdomen/Pelvis	CT Abdomen, CT Pelvis, CT Abdomen/Pelvis
Heart	SPECT
Knee	CT Knee
Shoulder	CT Shoulder
Sinus	CT Sinus
Spine	CT Lumbar Spine, MRI Lumbar Spine
Thorax	CT Thorax

After identifying duplicate records, we applied the following set of rules, sequentially, to consolidate the duplicate records into a single record:

3. If one or more records in a set of duplicates had been successfully matched to a claim, we retained all matched records and dropped the unmatched records from the analysis.
4. If no records were matched, we relied on termination codes that a convener added to a DSS record to designate a record as canceled. If the remaining records included both "terminated" and "open" records (where "open" denotes nonterminated and unmatched records), we retained the "open" records.
5. If all records in the duplicate set were either terminated or "open" we retained the record with the highest appropriateness score.

Measuring Changes and Cancellations to Orders

Changes to orders were measured in two ways. For all orders that were not associated with duplicate orders before our processing of duplicates (described above), we examined differences between the initial order and final order within the same DSS record. We defined a change as a difference between the initial order and final order in terms of (1) imaging procedure, (2) contrast instructions, or (3) study type (single or multiple imaging studies—which is relevant for SPECT orders only). For orders that were the result of processing duplicate orders, we identified a change as any difference in procedure, contrast, or SPECT study type *between* the multiple records of each set of duplicates or *within* any single record (i.e., between the initial order and final order) within a single DSS record.

Cancellations were measured directly by the DSS system and represent cancellations made by the ordering clinician immediately after receiving feedback. Cancellations made subsequent to a clinician's engaging with DSS, such as a cancellation to an order because of a patient's worsening condition) are not considered a cancellation for the purposes of these analyses.

Methods for Analyses of Imaging Utilization Using Claims Data

Measuring MID Imaging Procedure Utilization

We measured utilization at the ordering clinician level. In each month, we counted the number of imaging procedures ordered by each ordering clinician, as indicated by the referring provider field and date of service on carrier claims. We then created a rate of utilization of imaging procedures per unique Medicare beneficiary treated per ordering clinician per month. The denominator of this measure is the number of unique Medicare beneficiaries treated per physician per month (a beneficiary is identified as "treated" if they received a procedure or evaluation and management visit; procedures were identified as carrier claims with Berenson-Eggers Type of Service (BETOS) codes with first character "P" and visits were identified as carrier claims with BETOS codes with first character "M"). This measure adjusts utilization counts for the number of unique beneficiaries treated, a measure of clinical productivity for each clinician in each month.

We identified MID imaging procedures using Medicare claims files. We first identified all professional component and global service claims with MID imaging procedure codes (Table A.2) in the carrier file.

Table A.2. HCPCS Codes for MID Imaging Procedures

Demonstration Advanced Imaging Procedure	HCPCS Codes
CT Abdomen	74150, 74160, 74170
CT Pelvis	72192, 72193, 72194
CT Abdomen and Pelvis	74176, 74177, 74178*
CT Brain	70450, 70460, 70470
CT Lumbar Spine	72131, 72132, 72133
CT Sinus	70486, 70487, 70488
CT Thorax	71250, 71260, 71270
MRI Brain	70551, 70552, 70553
MRI Lumbar Spine	72148, 72149, 72158
MRI Knee	73721, 73722, 73723
MRI Shoulder	73221, 73222, 73223
SPECT MPI	78451, 78452

We then matched all professional component claims to technical component claims in the carrier or outpatient files. We used the following matching criteria:

- *Exact match:* For every imaging procedure technical component claim, identify a claim on the carrier file with any HCPCS modifier='26' (professional component) and matching the technical component claim as follows: same HCPCS code, same health insurance claim, date of service within the seven-day window following the technical component date of service.

- *Loose match:* For imaging procedure technical component claims without successful exact matches, check for "loose matches" with same criteria as above except HCPCS must be in same family (i.e., listed on same line of Table A.2). For example, if the technical component claim has HCPCS 74150 (CT abdomen without contrast), allow a match with carrier claim with HCPCS 74160 (CT abdomen with contrast) or 74170 (CT abdomen with and without contrast) and meeting other matching criteria.
- *Exclude overreads:* exclude multiple professional component claims that occur within a seven-day window of a matching technical component claim or a global service claim for the same patient.

We then applied two types of exclusion criteria to exclude imaging procedures performed in the inpatient or emergency department setting. First, we applied code-based exclusions using codes on both the professional and technical components of claims. Imaging procedure claims were excluded using the following criteria:

- Technical component, professional component, or global service carrier claim lists an excluded Line Place of Service Code (Table A.3).
- Outpatient technical component claim lists an excluded Facility Type Code (Table A.4), Service Classification Type Code (Table A.5), or Revenue Center Code (Table A.6).

Table A.3. Excluded Line Place of Service Codes

Excluded Codes
23: Emergency room—hospital
21: Inpatient hospital
20: Urgent care facility (eff. 1/1/03)
81: Independent laboratory
65: End stage renal disease treatment facility
31: Skilled nursing facility (SNF)
42: Ambulance—air or water
51: Inpatient psychiatric facility
61: Comprehensive inpatient rehabilitation facility

Table A.4. Excluded Facility Type Codes

Excluded Code
8: Special facility or Ambulatory Surgical Center (ASC) surgery
2: SNF
7: Clinic or hospital-based renal dialysis facility
3: Home health agency (HHA)
4: Religious nonmedical (hospital) (effective August 1, 2000)
5: Religious nonmedical (extended care) (effective August 1, 2000)
6: Intermediate care
9: Reserved

Table A.5. Excluded Service Classification Type Codes

Excluded Codes
5: Critical Access Hospital (effective October 1999) formerly rural primary care hospital (effective October 1994)
2: Hospital based or Inpatient (Part B only) or home health visits under Part B
4: Other (Part B)—(Includes HHA medical and other health services not under a plan of treatment, hospital or SNF for diagnostic clinical laboratory services for "nonpatients," and referenced diagnostic services.
2: Hospital-based or independent renal dialysis facility
3: Ambulatory surgical center in hospital outpatient department
1: Inpatient (including Part A)
1: Rural Health Clinic
1: Hospice (nonhospital-based)
2: Hospice (hospital-based)
3: Free-standing provider-based federally qualified health center (effective October 1991)
4: Other rehabilitation facility (ORF) and community mental health center (effective October 1991 through March 1997); ORF only (effective April 1997)
4: Freestanding birthing center
5: Intermediate care—level I
5: Comprehensive rehabilitation center
6: Intermediate care—level II
6: Community mental health center (effective April 1997)
6-8: Reserved for national use
7: Subacute inpatient (revenue code 019X required) (formerly Intermediate care—level III) NOTE: 17X and 27X are discontinued effective October 1, 2005.
7-8: Reserved for national assignment
8: Swing beds (used to indicate billing for SNF level of care in a hospital with an approved swing bed agreement)
9: Reserved for national assignment
9: Other

NOTE: A single Service Classification Type Code value has different interpretations based on the facility type code on the claim.

Table A.6. Excluded Revenue Center Codes on Outpatient MID Imaging Procedure Claims

Excluded Code	Revenue Center Label
0450	Emergency room-general classification

Second, we excluded imaging procedures with an emergency room visit or inpatient stay that occurred on the same day. Emergency room visits were defined as inpatient claim line items or outpatient claim line items with revenue code=0450–0459 (emergency room), 0981 (professional fees—emergency room) OR (revenue code=0760, 0762 [observation hours] AND HCPCS_CD="G0378"). Inpatient stays were defined as inpatient claims with admission date or discharge date not missing (or 00000000).

In order to further examine imaging utilization patterns, we also created measures of utilization of "non-MID imaging procedures," defined as imaging procedures not included in MID and identified as substitutes for MID imaging procedures by two primary care physicians on our team (Kahn, Wenger) and a radiologist consulting with our team (Lee).

Identifying Demonstration Group Ordering Clinicians

The "demonstration group" of ordering clinicians should include clinicians who are exposed to the intervention; that is, clinicians working at practices participating in the demonstration and using DSSs to order MID imaging procedures. We followed an intention-to-treat evaluation design whereby the demonstration group was identified based on the 18-month period preceding the initiation of the demonstration and then followed throughout the demonstration. This design was used to protect against attrition bias.

Our general approach to identifying the demonstration group was to first identify a "finder file" including all clinicians who could be part of the demonstration group, and then identifying the subset of clinicians from the finder file that comprise the demonstration group study sample for utilization analyses using claims data.[13]

We used three data sources to identify the demonstration group ordering clinician finder file:

- Claims—we identified clinicians associated with practices participating in the demonstration using the tax identification number (TIN) on claims, and we identified clinicians ordering MID imaging procedures (defined as described above).
- DSS records—only (and theoretically, all) clinicians associated with practices participating in the demonstration and ordering MID procedures have DSS records.
- Workbooks—conveners were asked to submit a list of all physicians and nonphysician practitioners who are authorized by state law to order advanced imaging services. However, not all of these clinicians actually order MID procedures. In preparing the workbooks, conveners encountered some difficulty in identifying all affiliated clinicians who could order MID procedures.

To identify ordering clinicians in claims for the finder file, we first identified all beneficiaries who received at least one service from a demonstration TIN in the 18 months prior to demonstration initiation (April 1, 2010–September 30, 2011), then identified all clinicians who ordered MID imaging procedures for those beneficiaries during the time period. This resulted in a group of 85,034 ordering clinicians. We identified the subset of these ordering clinicians who were affiliated with a demonstration TIN (billed using the TIN for evaluation and management visits, procedures, and/or imaging during the 18-month time period). This resulted in 6,498 ordering clinicians identified in claims and added to the demonstration group finder file.

The "DSS" group included all clinicians who appeared on DSS Submission 05 data files, which include orders from October, 1, 2011, until approximately November 30, 2012, with some variation by practice/convener. The "workbooks" group included all clinicians that were listed on the workbooks accompanying Submission 05. We excluded ordering clinicians from the

[13] The study sample of clinicians in the claims-based analyses does not exactly match the study sample for DSS-based analyses, which is limited to clinicians appearing in the DSS. The DSS-based analyses will also use the order rather than clinician as the unit of analysis.

finder file who ended demonstration participation before the demonstration baseline period began (n=603) or who joined the demonstration after the baseline period had initiated (n=423), as indicated in the workbooks.

We included the following subgroups from the finder file in the demonstration group study sample:

6. The 3,858 ordering clinicians identified in all three data sources: claims, DSS, and workbooks.
7. The 1,746 ordering clinicians identified in claims and workbooks but not DSS, and the 22 ordering clinicians identified in claims only. These clinicians had a claim for an ordered MID imaging procedure in the 18 months prior to the demonstration, but did not use DSS in the baseline period of the intervention. These clinicians may have not used DSS for several reasons, including: they did not have a reason to order an exam during the time period included in the DSS Submission 05 data; they use a proxy for DSS and the proxy is identified on the DSS record rather than the ordering clinician identified on claims; they wanted their patient to have an exam but chose for a different provider (e.g., a specialist from a different practice) to order the image; or they had access to methods of ordering that do not require DSS. In sensitivity analyses, we excluded these clinicians from the demonstration group.
8. The 118 ordering clinicians identified in claims and DSS but not workbooks. These clinicians ordered MID imaging procedures at demonstration practices but were not identified by conveners in the workbooks.

We excluded the following subgroups of finder file ordering clinicians from the demonstration group study sample:

9. The 1,135 clinicians that used the DSS in the baseline period but were not identified in claims (127 of these clinicians were also not listed in Workbooks). It is possible that we would not identify any MID imaging procedures ordered by these clinicians in claims, leading to a measured utilization rate of zero. Therefore, it is possible that we would undercount the utilization of MID imaging procedures if we included these clinicians. In contrast, the comparison group of ordering clinicians has been selected based on the fact that they appear as the ordering clinician on at least one MID imaging procedure claim. For this reason, we excluded clinicians that did not have any MID imaging procedures identified in claims from the study sample for analyses of claims-based outcomes (i.e., imaging costs and utilization). However, this group was used in analyses of DSS-based outcomes (e.g., appropriateness of orders) and could be used in sensitivity analyses for claims-based analyses. There are several reasons why these clinicians appeared in DSS but not claims, including:

a. They ordered MID imaging procedures in the baseline period but not the 18 months prior to the demonstration and therefore were not identified in claims.
b. The NPI entered on DSS records may not match the NPI listed on claims; for example, a resident physician NPI may be entered on DSS data and the attending physician NPI entered on claims, or the NPI on DSS data was for a nonphysician provider who is never listed as the ordering NPI on claims.
c. A DSS order was placed but the imaging procedure wasn't rendered during the DSS Submission 05 time period; possibly, the patient may have decided not to receive the imaging procedure after the order had been placed.
d. The DSS order was placed for a beneficiary not enrolled in Medicare fee-for-service and the record was included in the DSS data by error.
e. The clinician joined the demonstration practice after the baseline period started, but was not identified and excluded as a late entrant due to incomplete or incorrect information in the workbooks.

10. The 11,209 clinicians that appeared in workbooks only. This group could include clinicians affiliated with demonstration TINs and legally eligible to order MID imaging procedures, but who do not actually order MID imaging procedures.

Identifying Comparison Group Ordering Clinicians

A comparison group of ordering clinicians is necessary to examine trends in utilization over time relative to the demonstration group. To identify a comparison group of ordering clinicians, we first identified a sample frame including all clinicians practicing in counties that resemble counties in which demonstration clinicians practice. Using the practice site street address from the workbooks, we identified the complete set of counties participating in MID (i.e., "demonstration counties"). We then identified three matching comparison counties for every demonstration county.

County matching was performed with respect to eight characteristics: census region, an urban-rural indicator, number of generalists, number of specialists, number of inpatient days, number of outpatient visits, MID imaging utilization, and per capita income (Table A.7).

Table A.7. County-Level Characteristics Used to Match Control Counties to Intervention Counties

Characteristic	Variable definition	Source
Census region	Categorical: 1. Northeast 2. Midwest 3. South 4. West	Area Resource File (ARF)
Rural/urban	Categorical: 1. Metro areas of 250,000 population or more 2. Urban population of 2,500 or more, adjacent to a metro area 3. Completely rural or urban population of less than 20,000, not adjacent to a metro area	ARF
Total general practice physicians per capita	Numerator: Total number of general practice doctors in county (2008) Denominator: County population estimate (2008)	ARF
Total specialist physicians per capita	Numerator: Total number of specialist doctors in county (2008) Denominator: County population estimate (2008)	ARF
Inpatient days	Numerator: Total number of inpatient days (2007) Denominator: County population estimate (2008)	ARF
Outpatient visits	Numerator: Total number of outpatient visits in short-term general hospitals, short-term nongeneral hospitals, and long-term hospitals (2007) Denominator: County population estimate (2008)	ARF
Prebaseline per capita imaging utilization	Numerator: Count of technical component claims with a MID HCPCS code in 2009 for beneficiaries living in county X Denominator: Weighted count of beneficiaries in county X, defined as: Sum of Part B coverage months for beneficiaries in the 2009 5-percent beneficiary summary file, divided by 12	Medicare claims 5% sample
Per capita income	Per capita income (2007)	ARF

Since census region and the urban-rural indicator are categorical variables and were deemed important for obtaining good matches, we imposed that the control counties would match the demonstration counties exactly with respect to these two characteristics.

Counties were matched using the following algorithm. For any given demonstration county, we first selected the group of potential comparison counties for that particular demonstration county, including only those nondemonstration counties that match exactly the demonstration county with respect to census region and urban/rural category.

For each demonstration county, we calculated a "distance measure" to summarize the difference between the county and potential comparison counties on the six remaining variables. The distance measure was defined as the mean squared difference between a potential comparison county and the demonstration county:

$$\frac{\sum_{i=1}^{6}(X_{Ci} - X_{Di})^2}{6}$$

where *i* indicates the six matching variables, X_{Ci} is the value for potential comparison county *C* for variable *I*, and X_{Di} is the value for demonstration county *D* for variable *i.*

Since the six matching variables are all on different scales, we rescaled these variables so that they are on the same scale. As an illustration of the need for rescaling, assume that income per capita differs by $500 between two counties, while the number of specialists per capita differs by

0.001. If we did not rescale the variables, the county minimizing the sum of the mean squared differences of these two variables will inevitably be the one with the county with closest income per capita to the demonstration county; because income per capita is on a larger scale, it would have more weight. Therefore, we transformed each of the six variables so that they have mean 0 and standard deviation 1 within each group of counties in the same census region and with the same urban-rural category.

We then identified all clinicians practicing in comparison counties who were in the comparison group sample frame. We defined eligibility for the sample frame as NPIs with business location address ZIP code in NPPES located within a comparison group county, and within defined specialty categories "Allopathic and Osteopathic Physicians" or "Physician Assistants and Advanced Practice Nursing Providers." This resulted in 146,101 unique NPIs.

Next we identified the subset of these clinicians listed as the ordering clinician on MID imaging procedure claims in the 18 months prior to demonstration initiation (April 1, 2010, through September 30, 2011), limiting MID imaging procedure claims to those that passed both visit-based and code-based exclusions for procedures performed in the inpatient or emergency room setting. These restrictions resulted in 41,538 unique NPIs in the comparison group. Some of these clinicians are counted as comparisons for more than one convener because some comparison counties were matched to more than one demonstration county.

Propensity Score Weighting

We fit propensity score models separately for each convener using both the demonstration and comparison clinicians. We used a convener-specific approach because a single model may not allow high-quality matching of control clinicians to the unique profile of clinicians associated with each convener. The propensity score model was a logistic regression model in which the dependent variable was an indicator of whether the clinician was a participant in MID. The predicted probabilities from this model are known as propensity scores and indicate the probability each clinician was a participant in MID.

We developed the predictor variables used in the propensity score models using 100 percent claims files and the NPPES. We used the following variables:

- *Clinician characteristics.* We used 100-percent claims files to measure predemonstration MID ordering volume, predemonstration non-MID ordering volume, and patient volume. We measured clinician specialty using provider taxonomy codes on NPPES.
- *Beneficiary characteristics.* We aggregated beneficiary characteristics and included them as predictors in the clinician-level propensity score model. First, we identified all patients for which each clinician filed at least one claim. We then estimated the following characteristics:
  - **Mean HCC scores** were used to represent the expected yearly cost of a Medicare beneficiary based on his or her diagnoses and demographic information.

- **Standard deviation of HCC scores** for each clinician were included to account for differences across providers in the variability in case mix of the patients they treat.

We measured balance between the demonstration and comparison groups, calculating the standardized mean difference on the observed characteristics listed above. This measure is simply the absolute difference in means for each predemonstration covariate, across the treated and control designations, divided by a standard deviation in order to bring the differences into a comparable scale. Conventionally, standardized mean differences of 0.2 are considered "small," 0.5 is considered "moderate," and 0.8 is considered "large" (McCaffrey, Ridgeway, and Morral, 2004).

In Table A.8, we display the maximum (across covariates) of the standardized mean differences between the treated and control conditions. From this table, we see that prior to weighting, the imbalances for each convener is in the moderate-to-large range (from 0.39 to 1.04). After applying the propensity score weights, however, the standardized mean differences are all at or below the 0.2 cutoff for small differences, as is preferred in such analyses. Consequently, we have little concern that the variables used in the propensity score model (clinician's specialty; mean and standard deviation of HCC; and number of beneficiaries and MID and non-MID utilization during the predemonstration period) are confounding our results.

Table A.8. Maximum Standardized Mean Differences for Unweighted and Propensity-Score Weighed Controls

Convener	Unweighted	Propensity-Score Weighted
A	0.55	0.15
B	0.54	0.09
C	0.39	0.09
D	1.04	0.07
E	0.68	0.13
F	1.98	0.20
G	0.63	0.15

Regression Approach

We used a difference-in-difference analysis with a comparison group to measure the impact of the intervention on trends in utilization. The analysis isolated the effect of the intervention by comparing changes in the outcome (average number of MID images per Medicare beneficiary) in the intervention group to changes in the outcome of the comparison group. The main feature of this analysis is that it addresses two major threats to validity: (1) selection bias from the voluntary nature of the demonstration, and (2) confounding due to secular trends in imaging utilization independent of the demonstration. The analysis allows for changes in utilization rates for the control group as the demonstration passes from the predemonstration period through the baseline period and from the baseline phase to the demonstration period. By comparing the mean, provider-level differences in utilization between these periods for the demonstration and control providers, we are able to quantify the incremental increase or decrease in utilization levels for the demonstration group relative to the comparison group.

Technical Appendix B: Evaluation of the MID: Focus Group Methodology

As part of the evaluation of MID, RAND conducted a series of focus groups with physicians (including generalists and specialists) and staff from practices involved in the demonstration. These focus groups were supplemented with patient focus groups. The purpose of the focus groups was to collect information on clinician, staff, and patient experiences using ROE for advanced image ordering.

Clinician and Staff Focus Groups

Clinician and Staff Sample Selection

One responsibility of each MID convener was to generate continually updated workbooks listing all clinicians who could order advanced imaging for Medicare FFS beneficiaries from all practices associated with the convener. From this list, RAND identified a sample of candidate physicians for focus groups from the practices affiliated with each of the five conveners taking part in the MID demonstration. While quantitative analyses were conducted across seven conveners, the sampling plan and conduct of focus groups was already fixed by the time the large volume for some of the subconveners was noted. Accordingly, the clinician and staff focus group was conducted for the original five conveners, while DSS and claims analyses are conducted for seven conveners (including large subconveners).

In selecting a sample of candidates to include in the focus group, we aimed to include a mix of providers in terms of practice site and physicians' specialties, including generalists (internal medicine, family medicine, and geriatrics) and specialists (cardiology, gastroenterology, hematology, neurology, neurosurgery, urology, oncology, orthopedics, otolaryngology, pulmonary, rheumatology, and general surgery). We also aimed to include a balance of physicians in terms of practice location and volume of advanced imaging ordered during the year prior to MID initiation.

For each physician focus group, RAND identified a primary list of ten candidates and then a backup list of up to ten physicians per primary candidate to participate in the focus group in the event that a physician on the primary list of candidates turned out to be ineligible (for example, if a candidate were no longer at the practice), unavailable, or unwilling to take part in the focus group.

From the offices of clinicians who participated in MID focus groups, staff nurses, other clinical staff, and administrative assistants who ordered MID images or assisted clinician

participants in submitting advanced imaging orders using DSSs were invited to participate in staff focus groups.

Recruitment of Focus Group Participants

Prior to starting recruitment, RAND conducted a presentation (via webinar) attended by representatives from each of the five conveners and by CMS. The purpose of the presentation was to provide the conveners with an overview of the plan for conducting the focus groups, and to seek their cooperation in recruiting participants for the focus groups. During these presentations, RAND also described the approach to be used in recruiting physicians and staff for the focus groups.

The recruitment approach proposed by RAND involved sending each focus group candidate an advance notification letter printed on letterhead (which included the CMS, convener, and RAND logos) and signed by a CMS representative and by the principal investigators from RAND and from the convener site. RAND sent each convener a draft of the advance notification letter to be sent to focus group candidates, a series of frequently asked questions they could use in addressing issues raised by those contacted to participate in the focus group, and the list of candidates for the focus groups. Conveners were asked to review the lists to confirm that all candidates remained eligible for the focus groups and to provide candidates' contact information (mailing address, telephone and fax number, and email addresses).

RAND began recruiting focus group candidates in early spring 2012. Focus groups began in May 2012 and continued for an 18-month window, until September 2013. Although the same recruitment steps were used across conveners, there were important differences that facilitated recruitment at different sites. Some conveners opted to contact focus group candidates by email first, followed by the advance notification letter. While some sites preferred to mail the advance notification letter themselves, most preferred to have RAND mail the letter to focus group candidates directly. All but one of the conveners opted to have RAND conduct the bulk of the recruitment.

In recruiting physician candidates, RAND first contacted each candidate via email, and then followed up by phone as necessary to confirm participation. We first contacted the ten physicians prioritized for inclusion in the focus group, and only contacted physicians on the back-up list if the physicians on the priority list were not eligible, interested, or available to participate. Each focus group candidate had several back-ups that were matched based on specialty and volume of advanced imaging ordered. In recruiting a replacement, we went down the list of back-up physicians and aimed, as much as possible, to maintain a mix of doctors in terms of specialty, practice location, and volume of advanced imaging ordered.

Once a physician agreed to participate in a focus group, s/he received a confirmation email with instructions for taking part in the group and a PDF copy of a survey to complete and return prior to the start of the group. A few days before the date of the focus group, we sent a reminder email to all participants and also called to remind them again the day before each focus group.

Participants who had not completed the survey at the time of the focus group were encouraged to complete the survey during or after the meeting.

Focus Group Methods

An experienced moderator handled all the focus groups. In most cases, Dr. Kahn, the project's principal investigator, also helped moderate the focus groups. The moderator used the same scripted protocol across focus groups, although some of the topics were reorganized after the first two focus groups to improve the flow of the discussion. All focus groups were conducted via web-assisted conference call. Participants were asked to call into a toll-free number and log into a web page to view the focus group discussion questions. Participants were encouraged but not required to take part in the discussion and to identify themselves before responding to a question or making an observation. In addition to the moderators, a note-taker was present. All focus groups were audio-recorded and professionally transcribed. Physician participants were mailed a check for $150 to thank them for participating.

Table B.1 summarizes the results of the recruitment process. Overall, 65 clinician focus group participants, including 33 generalists, 32 specialists, and 27 staff completed the survey. Among clinician responders, the median age was 48 years, and the median time since medical school graduation was 21 years (range 16–31 years). Seventy-nine percent of respondents were involved with teaching medical students, residents, or fellows. In most instances, they taught multiple types of trainees. None of the respondents identified a solo practice as their main practice site. One-third of the respondents indicated they practiced in single specialty groups and two-thirds practiced in multispecialty groups.

Table B.1 Focus Group Recruitment

	Generalists	Specialist	Staff[c]	Generalist	Specialist	Staff[b]	Generalist	Specialist	Staff[c]	Generalist	Specialist	Staff[c]	Generalist	Specialist	Staff[c]	Participants
Conveners		A			B			C			D			E		ALL
Total Contacted[a]	—	—	—	23	36	—	75	65	—	39	41	—	98	41	—	418[d]
Agreed to participate	5	7	3	5	6	5	15	6	—	9	12	—	9	8	9	99
Participated	5	7	3	4	6	5	13	5	—	8	11	11	5	6	8	97
Completed Survey	5	5	3	4	7	6	9	4	—	7	10	11	8	6	7	92

[a] Includes clinician and staff members approached about possible focus group participation

[b] Convener C staff did not participate in focus groups

[c] Staff using MID or helping clinicians use MID in practices where clinicians were involved with focus groups were invited to participate in staff focus groups

[d] Convener A conducted its own recruitment so contact data are not available.

In general, the study was designed with descriptive rather than inferential goals in mind. That is to say, the study sought to elicit opinions from a broadly representative collection of clinicians in terms of specialty, order volume, and physician age for each convener. In this way, the study design provides our best effort at producing a snapshot of clinician attitudes toward the intervention given the study's resource constraints.

While such an approach guarantees wide coverage of important classifications of physicians participating in the evaluation, the downside of such a broadly representative approach is that it compromises our ability to perform statistical inference, such as computing confidence intervals for quantities such as average physician satisfaction with DSSs. For example, in many cases, a single physician was selected from each of several specialties within a convener. Sampling a single unit from a prespecified group of individuals (such as a specialty/convener pair) precludes the possibility of estimating variation within that group, which, in turn, affects our ability to calculate the variance components needed for statistical hypothesis tests and confidence intervals. While our conclusions will therefore not be phrased in statistical terms, we feel that the broad agreement of clinicians across specialties and providers on many of the survey questions constitutes strong evidence for many of our conclusions.

Thematic Analysis of Focus Group Data

The focus group transcripts were reviewed by the moderator(s) to ensure accuracy. Audio-recordings were consulted to fill in any gaps. An experienced qualitative researcher analyzed focus group transcripts to identify key themes. The process involved coding focus group transcripts using MAXQDA 10, a qualitative data analysis software that marks blocks of text pertinent to specific themes. The marked text is then divided into shorter quotations that represent an idea and can be read as independent statements.

A hierarchically organized codebook was developed based on the focus group protocol to group coded text into themes (MacQueen et al., 1998). The codebook facilitates data coding, helps to ensure coding consistency across focus groups, and ensures the production of comparable information based on each focus group transcript. Analysis of focus group data started after the first focus group was conducted; the codebook was refined after each subsequent focus group to improve the data coding approach based on new information that emerged from each focus group and to identify any emergent themes.

We used both theory-driven (e.g., deductive) and data-driven (e.g., inductive) approaches to thematic data coding to describe the process of advanced image ordering, as well as the process and perceived outcomes of the MID project (Fereday and Muir-Cochrane, 2008). For example, because research shows that effective implementation is associated with better intervention outcomes and implementation itself depends on the extent to which education and technical assistance were provided, we were interested in understanding how the MID intervention was implemented by each convener and whether clinicians and staff received training and technical assistance (Durlak and DuPre, 2008). Therefore, we created codes that describe different aspects

of MID implementation based on the existing literature before we started coding focus group transcripts. Moreover, we also coded transcripts inductively to identify any unanticipated themes and topics, such as overriding orders that have been rated equivocal or inappropriate by the decision support system. Reliance on both deductive and inductive approaches to coding qualitative data helped us ensure that similar themes were abstracted for each convener and focus group type (generalist, specialist, staff), which is important for synthesizing information across different focus groups and identifying overarching findings.

Once all focus group transcripts were coded, a version of the constant comparative method of qualitative analysis was used by a qualitative expert to identify any differences within and between conveners and focus group types in terms of their advanced image ordering process, experiences with ROE and DSS, perceptions of the potential impact on the MID demonstration on advanced image ordering appropriateness, utilization, and costs (MacQueen et al., 1998; Fereday and Muir-Cochrane, 2008; Boeije, 2002; Durlak and DuPre, 2008; Glaser 1965). By using a comparative approach to qualitative data analysis, we were able to identify the aspects of this demonstration project that were: (1) perceived by clinicians to be qualitatively associated with perceived effectiveness and success of DSSs; (2) recurrent across conveners and focus group types; and/or (3) unique to generalists or specialists.

Brief Survey of Focus Group Participants

Clinicians and staff who agreed to participate in focus groups were invited to complete a brief survey designed to supplement the focus groups, available as either a paper-based or web-based document. While focus group items were designed to motivate discussion and prompt group interaction, survey items were developed to learn about individual participants' experiences with various aspects of MID. Table B.1 documents the number of survey respondents. Key findings from the survey respondents are embedded in the focus group findings.

Patient Focus Groups

A similar methodology was applied for patient focus groups. To collect information on the impact of the MID on patient satisfaction and patients' experience of care, we conducted two on-site focus groups with patients during February 2014. In aggregate, analyses from these patient focus group and from analyses of physician and staff focus groups regarding their perception of patients' experiences of advanced imaging and decision support informed RAND's perspective on patient's experience with MID.

In order to be eligible for the focus groups, participants had to be adult Medicare beneficiaries who had had one of the 12 advanced images included in MID in the last three to six months. In addition, candidates for the focus groups had to speak English well enough to

participate in the focus group, had to be able to attend the focus group in person, and had to be able to provide informed consent.

Once eligible patients agreed to participate in a focus group, they received instructions for taking part in the focus group (directions to the focus group location, parking instructions, date and time of the focus group). In addition, we called participants to remind them of the focus group the day before the event. A total of 13 patients participated in the focus groups (7 in the morning focus group and 6 in an afternoon focus group).

An experienced moderator conducted both focus groups in person, using the same scripted protocol. In addition to the focus group moderator, a RAND note-taker was present. Site-level staff members were not allowed to take part in or observe the focus groups. The focus groups were audio-recorded and transcribed. Participants were asked to review and sign a consent form before the start of the focus group. In addition, they were asked to complete a brief, one-page survey that collected basic demographic information. A light meal was served at the beginning of the meetings. We gave each participant $100 in cash at the end of the focus group to thank them for their time.

Professionally transcribed focus group transcripts were reviewed to ensure accuracy. Audio-recordings were consulted as necessary to fill in any gaps in the transcripts. The focus group transcripts and notes were reviewed and analyzed thematically (using the main topics and questions covered in the focus groups), to identify key findings. A second review of the audio recordings and focus group transcripts was conducted in order to identify verbatim quotes that reinforce key findings.