Medicare Imaging Demonstration Evaluation Report to Congress

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 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 to 16.1 percent and rates of inappropriate ordering between 5.3 percent to 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.