Comparison of Content on the American Board of Internal Medicine Maintenance of Certification Examination With Conditions Seen in Practice by General Internists

Bradley Gray, PhD; Jonathan Vandergrift, MS; Rebecca S. Lipner, PhD; Marianne M. Green, MD

IMPORTANCE Success on the internal medicine (IM) examination is a central requirement of the American Board of Internal Medicine’s (ABIM’s) Maintenance of Certification program (MOC). Therefore, it is important to understand the degree to which this examination reflects conditions seen in practice, one dimension of content validity, which focuses on the match between content in the discipline and the topics on the examination questions.

OBJECTIVE To assess whether the frequency of questions on IM-MOC examinations were concordant with the frequency of conditions seen in practice.

DESIGN, SETTING, AND PARTICIPANTS The 2010-2013 IM-MOC examinations were used to calculate the percentage of questions for 186 medical condition categories from the examination blueprint, which balances examination content by considering importance and frequency of conditions seen in practice. Nationally representative estimates of conditions seen in practice by general internists were estimated from the primary diagnosis for 13,832 office visits (2010-2013 National Ambulatory Medical Care Surveys) and 108,472 hospital stays (2010 National Hospital Discharge Survey).

EXPOSURES Prevalence of conditions included on the IM-MOC examination questions.

MAIN OUTCOMES AND MEASURES The outcome measure was the concordance between the percentages of IM-MOC examination questions and the percentages of conditions seen in practice during either office visits or hospital stays for each of 186 condition categories (eg, diabetes mellitus, ischemic heart disease, liver disease). The concordance thresholds were 0.5 SD of the weighted mean percentages of the applicable 186 conditions seen in practice (0.74% for office visits; 0.51% for hospital stays). If the absolute differences between the percentages of examination questions and the percentages of office visit conditions or hospital stay conditions seen were less than the applicable concordance threshold, then the condition category was judged to be concordant.

RESULTS During the 2010-2013 IM-MOC examination periods, 3,600 questions (180 questions per examination form) were administered and 3,461 questions (96.1%) were mapped into the 186 study conditions (mean, 18.6 questions per condition). Comparison of the percentages of 186 categories of medical conditions seen in 13,832 office visits and 108,472 hospital stays with the percentages of 3,461 questions on IM-MOC examinations revealed that 2,389 examination questions (69.0%; 95% CI, 67.5%-70.6% involving 158 conditions) were categorized as concordant. For concordance between questions and office visits only, 2,010 questions (58.08%; 95% CI, 56.43%-59.72% of all examination questions) involving 145 conditions were categorized as concordant. For concordance between questions and hospital stays only, 1,456 questions (42.07%; 95% CI, 40.42%-43.71% of all examination questions) involving 122 conditions were categorized as concordant.

CONCLUSIONS AND RELEVANCE Among questions on IM-MOC examinations from 2010-2013, 69% were concordant with conditions seen in general internal medicine practices, although some areas of discordance were identified.
The purpose of the American Board of Internal Medicine’s (ABIM) Internal Medicine Maintenance of Certification examination (IM-MOC) is to assess the knowledge, diagnostic reasoning, and clinical judgment expected of a general internist who delivers care in office and hospital settings. Although there are many forms of quality and performance feedback, examinations such as the ABIM IM-MOC examination attempt to address an important gap in these measures by assessing an individual physician’s competence in medical knowledge and clinical judgment. Quality measures based on patient data cover only a portion of the conditions seen by physicians and generally do not address rare but serious conditions or adequately consider quality issues around diagnostic reasoning. Therefore, patient data may be insufficient to determine whether a physician can diagnose conditions that cover the breadth of the field or whether a physician can make the right judgment about appropriate care.1

Recent critiques of the ABIM IM-MOC examination suggest that physician practices have evolved over time and that the questions that comprise the IM-MOC examination do not reflect these changes in terms of their relationship to the medical conditions encountered by practicing general internists.2 Thus, considering that 91% of general internists are certified by the ABIM, measuring the degree to which examination content aligns with clinical practice is both a key element of the validity of the examination results and an important policy question.3

Examination validity criteria include content validity, which addresses the match between content in the discipline and the examination question topic; construct validity, which focuses on the theoretical construct being tested; and predictive and consequential validity, which refers to how well examination performance predicts physician quality.4,5 The objective of this study was to measure the content validity of the IM-MOC examination by assessing the concordance between the frequency of questions related to specific conditions on the 2010-2013 IM-MOC examinations and the frequency with which these conditions were seen in general internal medicine practice.

Methods

Frequency of Conditions Seen in Practice

Chart abstraction data from the National Ambulatory Medical Care Survey (NAMCS), conducted by the National Center for Health Statistics (NCHS), were used to construct national estimates of the frequencies of conditions seen during general internist office visits from 2010 to 2013. Estimates were drawn from 2-week samples of visits to office-based nonspecializing internists, referred to as general internists.6 Visit conditions seen were recorded by the NAMCS participating physician as primary diagnosis International Classification of Diseases, Ninth Revision (ICD-9) codes abstracted from medical charts. NAMCS visit-level population weights were used to estimate the population-weighted frequency of conditions seen by office-based general internists nationally. A description of the NAMCS methods is available from the NCHS.7-10

Hospital stay discharge note abstraction data from the National Hospital Discharge Survey (NHDS), conducted by the NCHS, were used to construct national estimates of the frequencies of conditions during hospital stays in 2010 (the most recent year available). Estimates were drawn from a sample of hospital stays excluding discharges following births and for patients younger than 18 years. Hospital stay conditions seen were based on primary diagnosis ICD-9 codes based on the discharge notes. National Hospital Discharge Survey hospital discharge population weights were used to estimate the population-weighted frequency of hospital stay conditions seen for patients aged 18 years or older who were not admitted to the hospital for a delivery. Hospital stays for younger patients and for deliveries were excluded because they occur in hospital setting in which general internists are unlikely to be attending. A description of the NHDS methods is available from the NCHS.11

This sample of hospital stays was chosen because there was no record of specialty types for treating physicians in the NHDS data set, and it was more likely that internists providing care in the inpatient hospital setting would care for this sample of patients at some point during their stay. For example, if a patient was admitted for an amputation because of a diabetes complication, a general internist may in part be responsible for caring for that patient as an attending physician, even though the internist did not perform the surgery.

Examination Question Condition Frequency

Questions from the 2010-2013 IM-MOC examinations were used as the source of question condition frequency. The related conditions for each question were based on categorization of the question by the test-writing committee in the examination blueprint. Examination blueprints are specification tables describing an examination’s scope of content, clinical tasks (eg, diagnosis, treatment), and functional requirements (eg, item and examination statistics). The purpose of the blueprint is to drive examination construction and to provide test takers insights into examination content.12 Because there were no “field” questions on the IM-MOC examination, all questions were categorized.

The blueprint structure is a hierarchical content framework designed to ensure that the questions selected for an examination represent the appropriate scope and frequency of
conditions and clinical tasks for a particular discipline. Within the IM-MOC blueprint, each question is assigned a primary medical content category (level 1; eg, endocrinology, cardiovascular disease) and a specific topic area or medical condition within the primary category (level 2; eg, diabetes, hypertension). The IM-MOC blueprint level 2 covers 187 unique conditions. These conditions excluded topic areas that are included in the blueprint, such as ethics, epidemiology, or prevention and screening, which do not refer directly to specific medical conditions.

Examination question frequency was linked to level 2 conditions seen in office visits and hospital stays through ICD-9 codes that reflect condition category descriptions (see eTable 1 in the Supplement for condition ICD-9 codes). Two conditions, small intestinal disease and colonic and anorectal disease, could not be distinguished using ICD-9 codes and were combined, resulting in 186 distinct condition categories. Examination condition categories were assigned ICD-9 codes using an iterative process that involved reviewing the full ICD-9 code set and assigning codes that were consistent with the blueprint level 2 condition descriptors.

For each of 186 condition categories, the population-weighted percentages of office visit conditions seen were calculated by dividing the population-weighted frequency of office visit conditions seen in a condition category by the total population-weighted frequency of office visits across all 186 condition categories. Hospital stay conditions seen were calculated in the same way. For each of the 186 condition categories, the percentages of examination questions were calculated by dividing the number of questions in a condition category by the total number of questions across all 186 condition categories.

The NAMCS and NHDS protocols have been approved by the NCHS Research Ethics Review Board, and these data are publicly available. The ABIM provided approval for use of examination question blueprint category data applied in the analysis.

Concordance Criteria
The study objective was to measure one dimension of content validity by evaluating whether the question percentage for each of the 186 condition categories was different than the corresponding percentage of conditions seen. To accomplish this, 2 differences were calculated for each condition category as the basis for concordance categorizations: (1) the difference between the percentage of questions for a condition category and the percentage of office visit conditions seen and (2) the difference between the percentage of questions for a condition category and the percentage of hospital stay conditions seen.

For office visits, the concordance difference threshold used to evaluate these differences was 0.5 SD of the weighted mean percentages (ie, population-weighted percentages) across the 186 office visit conditions seen. The same method was used to calculate the concordance threshold for hospital stays. The thresholds were based on the conditions seen because the goal of the analysis was to evaluate whether the examination emphasis (reflected in the frequency of questions by condition) was noticeably different than the frequency of conditions physicians see in practice. The 0.5-SD criterion was selected because it corresponds with the medium effect size proposed by Cohen as being large enough to be perceivable. Furthermore, the literature review by Norman et al reports that across numerous health-related studies, 0.5 SD was a common criterion for detectable differences and suggests that this stems from the natural ability of individuals to judge differences.

A condition category was judged to be concordant if the difference between examination question percentage and either the percentage of office visit conditions seen or the percentage of hospital stay conditions seen was less than the applicable concordance threshold in absolute terms. If this standard was not met, the condition category was judged to be discordant. Overall examination concordance was calculated by summing the applicable question percentages for concordant categories. In addition, examination concordance was also calculated separately for office visit conditions seen and hospital stay conditions seen by comparing the applicable absolute difference between percentages of conditions seen with the applicable threshold.

Sensitivity Analyses
Five separate concordance sensitivity analyses were conducted (eAppendix in the Supplement). The first 2 sensitivities considered the upper and lower 95% confidence bounds of the 0.5-SD thresholds. The third sensitivity considered that some physicians see patients only in the office setting by applying a discordance criterion that weighted office visits more than hospital stays. The fourth sensitivity addressed including only 2010 hospital data. The fifth sensitivity considered excluding stays involving surgical procedures when calculating hospital stay frequencies.

Statistical analyses were performed with SAS version 9.3 (SAS Institute Inc). Tests were 2-sided and \( P<.05 \) denoted statistical significance.

Results
Study Sample
The applicable 2010-2013 NAMCS sample included 13,842 office visits to 621 office-based general internists for which the conditions seen were mapped into the 186 blueprint condition categories. Weighted to population level, these data represent 85.9% of office visits to 87.3% of office-based general internists. Of the 14.1% of visits for which the primary diagnosis could not be mapped to level 2 blueprint condition categories, 6.7% were well visits or routine follow-up visits and 7.4% had an ICD-9 primary diagnosis that was not specific enough or was outside of the examination blueprint (eg, unspecified fatigue). The 2010 NHDS hospital sample included 108,472 hospital stays that were mapped into the 186 blueprint condition categories. Weighted to the population level, this represents 83.1% of all applicable hospital stays. The 16.9% of applicable hospital stays that could not be mapped represent ICD-9 codes outside the blueprint or were not specific enough.
During the 2010-2013 IM-MOC examination periods, 3600 questions (180 questions per examination form) were administered. Of these, 3461 questions (96.1%) were mapped into the 186 study conditions (mean, 18.6 questions per condition). Of the 139 unmapped questions (3.9%), most involved epidemiology (48 questions [34.5%]) and ethics (44 questions [31.6%]).

**Concordant Percentages of Questions and Conditions Seen**

The 0.5-SD concordance thresholds for the percentages of the 186 conditions seen (condition population-weighted mean percentage as described in the Methods section) was 0.74% for percentages of office visit conditions seen and 0.51% for percentages of hospital stay conditions seen. Applying these thresholds to the 3461 examination questions, 2389 questions (69.0%; 95% CI, 67.5%-70.6%) were categorized as concordant (ie, 158 of 186 condition categories were such that the difference between the percentage of questions and either the percentage of office visit conditions seen or the percentage of hospital stay conditions seen was less than the corresponding threshold) (eTable 2 in the Supplement). For example, thyroid disorders was classified as concordant because the absolute difference of −0.09% (95% CI of difference CI, −0.56% to 0.37%) between this condition's question percentage of 1.30% (95% CI of difference CI, 0.92%-1.68%) and office visits percentage of 1.39% (95% CI, 1.15%-1.69%) was less than the concordance threshold in absolute terms for office visits of 0.74%. For this condition, the difference between the question percentage and hospital stay percentage of 1.02% (95% CI, 0.64%-1.40%) was greater than the hospital stays concordance threshold of 0.51%. However, the overall concordance criterion that the absolute difference between the question percentage and the percentage of either office visit or hospital stay conditions seen was less than the corresponding threshold was met.

For concordance between questions and office visits only, 2010 questions (58.08%; 95% CI, 56.43%-59.72% of all examination questions) involving 145 conditions were categorized as concordant. For concordance between questions and hospital stays only, 1456 questions (42.07% of all examination questions; 95% CI, 40.42%-43.71%) involving 122 conditions were categorized as concordant (eTable 3 in the Supplement).

**Discordant Percentages of Questions and Conditions Seen**

A total of 1072 questions (30.97%; 95% CI, 29.43%-32.51% of all examination questions) involving 28 condition categories were categorized as discordant (the Table lists the percentages of questions, office visits, and hospital stays and differences for each of these conditions). Depending on the nature of the discordance, concordance could be reached for these condition categories by having either greater or fewer examination questions related to this condition.

For 708 of the discordant questions (20.5% of all examination questions; 95% CI, 19.1%-21.8%) involving 17 conditions, the question percentage was greater than the percentages of both the office visit and hospital stay conditions seen (ie, concordance could be reached by reducing the number of questions). The most prevalent of these conditions in terms of number of questions was liver disease, with 72 questions. For this condition, the question percentage of 2.08% (95% CI, 2.70%-2.56%) was greater than the office visit percentage of 0.28% (95% CI, 0.16%-0.47%) by 1.81% (95% CI, 1.31%-2.30%), which exceeded the concordance threshold for office visit conditions seen of 0.74%. For this condition, the question percentage was also greater than the hospital stay percentage of 0.64% (95% CI, 0.57%-0.72%) by 1.44% (95% CI, 0.96%-1.92%), which was more than the concordance threshold for hospital stay conditions seen of 0.51%.

For 161 of the discordant questions (4.65% of all examination questions; 95% CI, 3.95%-5.31%) involving 6 conditions, the question percentage was less than the percentages for both office visits (18.57%; 95% CI, 17.32%-19.88%) and hospital stays (15.98%; 95% CI, 15.63%-16.34%) (ie, concordance could be reached by increasing the number of questions). The most prevalent of these conditions in terms of number of questions was lower respiratory tract infections, with 66 questions, for which the question percentage of 1.91% (95% CI, 1.45%-2.36%) was less than the office visit percentage of 2.97% (95% CI, 2.41%-3.65%) as well as the hospital stay percentage of 3.94% (95% CI, 3.75%-4.12%).

For 203 of the discordant questions (5.87% of all examinations questions; 95% CI, 5.08%-6.65%) involving 5 conditions, the question percentage was between the office visit and hospital stay percentages (ie, concordance could be reached by either lowering or increasing the number of questions). The most prevalent of these conditions in terms of number of questions was dysrhythmias conduction defects, with 71 questions, for which the question percentage of 2.05% (95% CI, 1.58%-2.52%) was between the office visit percentage of 1.14% (95% CI, 0.88%-1.47%) and the hospital stay percentage of 3.05% (95% CI, 2.89%-3.22%).

The Figure shows the percentages of office visit conditions seen, hospital stay conditions, and question percentages for each condition category, ordered from low to high, and their concordance designations. The Figure shows that office visits and, to a lesser extent, hospital stays were more concentrated on the right side of the distribution than were examination questions, which were more equally distributed (ie, the distribution of conditions seen was more skewed than the distribution of examination questions).

The results of sensitivity analyses are shown in eTable 4 in the Supplement. These sensitivity results ranged from an increase in the percentage of questions in the discordant category of 3.0 percentage points (95% CI, 0.9%-5.2%), from 31.0% to 3.4%, when the care setting was considered to a decrease of 5.7 percentage points (95% CI, 3.6%-7.8%), from 31.0% to 25.3%, in the discordant category when the upper 95% confidence bound of the 0.5 SD of the conditions-seen percentage was applied as a concordance threshold.

**Discussion**

In this study comparing the percentages of 186 categories of medical conditions seen by general internists in office visits and hospital stays with the percentages of 3461 questions on
<table>
<thead>
<tr>
<th>Conditions</th>
<th>Questions % (95% CI)</th>
<th>Office Visits % (95% CI)</th>
<th>Hospital Stays % (95% CI)</th>
<th>Questions Minus Office Visits</th>
<th>Questions Minus Hospital Stays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liv</td>
<td>2.08 (1.60-2.56)</td>
<td>0.28 (0.16-0.47)</td>
<td>0.64 (0.57-0.72)</td>
<td>1.81 (1.31 to 2.30)</td>
<td>1.44 (0.96 to 1.92)</td>
</tr>
<tr>
<td>Ematologic malignancies</td>
<td>1.96 (1.50-2.43)</td>
<td>0.37 (0.20-0.66)</td>
<td>0.39 (0.33-0.45)</td>
<td>1.60 (1.09 to 2.11)</td>
<td>1.58 (1.11 to 2.04)</td>
</tr>
<tr>
<td>Disorders of calcium metabolism and bone</td>
<td>0.73 (1.30-2.17)</td>
<td>0.25 (0.14-0.43)</td>
<td>0.08 (0.05-0.11)</td>
<td>1.49 (1.03 to 1.94)</td>
<td>1.65 (1.22 to 2.09)</td>
</tr>
<tr>
<td>Interstitial lung disease</td>
<td>1.53 (1.12-1.94)</td>
<td>0.09 (0.03-0.28)</td>
<td>0.19 (0.15-0.23)</td>
<td>1.45 (1.02 to 1.87)</td>
<td>1.34 (0.93 to 1.75)</td>
</tr>
<tr>
<td>Valvular heart disease</td>
<td>1.39 (1.00-1.78)</td>
<td>0.20 (0.11-0.36)</td>
<td>0.35 (0.29-0.40)</td>
<td>1.19 (0.78 to 1.60)</td>
<td>1.04 (0.65 to 1.43)</td>
</tr>
<tr>
<td>Pericardial disease</td>
<td>1.07 (0.73-1.41)</td>
<td>0.00 (0.00-0.02)</td>
<td>0.05 (0.03-0.07)</td>
<td>1.07 (0.72 to 1.41)</td>
<td>1.02 (0.67 to 1.36)</td>
</tr>
<tr>
<td>Hemolytic anemias</td>
<td>1.07 (0.73-1.41)</td>
<td>0.00 (0.00-0.03)</td>
<td>0.04 (0.02-0.06)</td>
<td>1.07 (0.72 to 1.41)</td>
<td>1.03 (0.68 to 1.37)</td>
</tr>
<tr>
<td>Endocarditis and other cardiovascular infections</td>
<td>1.07 (0.73-1.41)</td>
<td>0.00 (0.00-0.03)</td>
<td>0.13 (0.10-0.17)</td>
<td>1.06 (0.72 to 1.41)</td>
<td>0.94 (0.59 to 1.28)</td>
</tr>
<tr>
<td>Disroders of cerebral function</td>
<td>1.07 (0.73-1.41)</td>
<td>0.28 (0.19-0.41)</td>
<td>0.46 (0.40-0.53)</td>
<td>0.79 (0.43 to 1.15)</td>
<td>0.61 (0.26 to 0.95)</td>
</tr>
<tr>
<td>Vasculitis</td>
<td>1.04 (0.70-1.38)</td>
<td>0.01 (0.00-0.03)</td>
<td>0.04 (0.02-0.06)</td>
<td>1.03 (0.69 to 1.37)</td>
<td>1.00 (0.66 to 1.33)</td>
</tr>
<tr>
<td>Platelet disorders</td>
<td>1.04 (0.70-1.38)</td>
<td>0.11 (0.05-0.26)</td>
<td>0.12 (0.09-0.15)</td>
<td>0.93 (0.58 to 1.28)</td>
<td>0.92 (0.58 to 1.26)</td>
</tr>
<tr>
<td>Congenital heart disease in adults</td>
<td>1.01 (0.68-1.34)</td>
<td>0.05 (0.01-0.15)</td>
<td>0.09 (0.06-0.12)</td>
<td>0.97 (0.63 to 1.30)</td>
<td>0.92 (0.58 to 1.25)</td>
</tr>
<tr>
<td>Central nervous system infections</td>
<td>0.92 (0.61-1.24)</td>
<td>0.01 (0.00-0.08)</td>
<td>0.16 (0.12-0.20)</td>
<td>0.91 (0.59 to 1.23)</td>
<td>0.76 (0.44 to 1.08)</td>
</tr>
<tr>
<td>Miscellaneous oncology</td>
<td>0.90 (0.58-1.21)</td>
<td>0.07 (0.03-0.16)</td>
<td>0.15 (0.11-0.19)</td>
<td>0.83 (0.51 to 1.15)</td>
<td>0.74 (0.43 to 1.06)</td>
</tr>
<tr>
<td>Coagulation factor disorders and thrombotic disorders</td>
<td>0.90 (0.58-1.21)</td>
<td>0.10 (0.04-0.26)</td>
<td>0.02 (0.01-0.03)</td>
<td>0.80 (0.47 to 1.13)</td>
<td>0.88 (0.56 to 1.19)</td>
</tr>
<tr>
<td>Adrenal disorders</td>
<td>0.84 (0.53-1.14)</td>
<td>0.01 (0.00-0.07)</td>
<td>0.05 (0.03-0.08)</td>
<td>0.82 (0.52 to 1.13)</td>
<td>0.78 (0.48 to 1.09)</td>
</tr>
<tr>
<td>Systemic sclerosis</td>
<td>0.84 (0.53-1.14)</td>
<td>0.03 (0.01-0.20)</td>
<td>0.01 (0.00-0.02)</td>
<td>0.76 (0.44 to 1.07)</td>
<td>0.83 (0.52 to 1.13)</td>
</tr>
<tr>
<td>Total</td>
<td>20.46 (19.15-21.77)</td>
<td>1.90 (1.52-2.38)</td>
<td>3.00 (2.83-3.16)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The difference between question and office visit percentages and the difference between question and hospital stay percentages are both positive or negative, the absolute difference between question and office visit percentages is greater than or equal to 0.74 (0.5 SD of visit percentage), and the absolute difference between question and discharge percentages is greater than or equal to 0.51 (0.5 SD of hospital stay percentage).

*Question percentage based on 3461 questions from examinations administered from 2010 to 2013.

Office visit percentage based on 2010-2013 National Ambulatory Medical Care Survey data applying nationally representative weights.

Hospital stay percentage based on 2010 National Hospital Discharge Survey data applying nationally representative weights.
Figure. Distribution of Percentages of Office Visits, Hospital Stays, and Questions for Each of 186 Conditions Seen

The top panel shows office visit percentages for each of 186 conditions, ordered from lowest to highest. The middle and bottom panels show percentages for each of 186 conditions for hospital stays and questions, respectively. Insets for each panel show the distribution among condition categories with less than 0.5% of office visit, hospital stay, or question percentages. The figure demonstrates that the distribution of conditions seen was more skewed than the question distribution.
IM-MOC examinations from 2010 to 2013, 69.0% of examination questions were concordant with conditions seen. This finding indicates that the IM-MOC examination has generally been consistent with the conditions seen in practice. However, with 31.0% of examination questions categorized as discordant, the study also identified potential opportunities for improvement.

The approach taken in establishing the concordance thresholds and applying these thresholds to 2 care settings should be considered in evaluating these results. The threshold for concordance was selected based on recommendations from the literature in terms of ability to perceive differences and for this study was selected as 0.5 SD of the weighted mean percentages of the corresponding 186 conditions seen. Yet, it is reasonable to consider other approaches for deriving a concordance threshold, perhaps based on a physician survey. As such, sensitivity analyses considered more and less stringent concordance thresholds and demonstrated that the main results were robust to several definitions and thresholds for concordance.

With respect to analyzing concordance across 2 settings, the criteria applied to construct concordance categorizes assumed that questions, absent a code for care setting, were applicable to either outpatient or inpatient settings. Although this assumption seemed reasonable overall, its application led to some counterintuitive concordance categorizations. For example, the question percentage for hypertension was judged to be discordant because it was similar to the hospital stay percentage (1.91% vs 1.84%), even though the question percentage was much lower than the office visit percentage for this condition (1.91% vs 13.87%). A care setting sensitivity analysis considered that this condition was highly discordant with office visit conditions seen yet was concordant with hospital stay conditions seen. Additionally, the data used to construct measures of concordance have been included in the eAppendix in the Supplement to enable evaluation of other measures of concordance.

A study finding related to the 2 settings was that IM-MOC examinations have been more concordant with conditions seen in office settings than in hospital settings. This may reflect that the examination is for general internists who practice in an outpatient setting or both outpatient and inpatient settings. The ABIM offers a separate hospital medicine examination for general internists who practice only in inpatient settings.

This study had several limitations. First, errors in diagnosis reflected in the medical record could have led to measurement error regarding conditions seen. This may be particularly problematic for less common conditions for which diagnostic experience could be lacking. Second, the NHDS does not link hospital stays to specific physician specialties. Because of this, the analysis assumed that general internists were involved at least in part in caring for all patients with hospital stays, with the exception of hospital stays for delivery and care for patient younger than 18 years. Third, the NHDS covered only one-quarter of the study period (2010), but those data were the most current data available at the time of the study. An analysis (eTable 4 in the Supplement) evaluating NHDS data from 2007-2010 suggests that using data from 2010 did not materially affect the study results as the distribution of hospital stay diagnoses appeared fairly stable in the years leading up to the 2010-2013 study period. Fourth, some of the concordant condition categories were fairly common in terms of conditions seen (eg, diabetes) but might have questions related to specific issues that are rarely seen (eg, an uncommon complication of diabetes).

Fifth, the study does not reflect changes to the blueprint that occurred in 2015. The analysis applied in this study could not yet be replicated with new data or the enhanced examination blueprint because more recent data on office visit conditions seen and hospital stay conditions seen are not currently available and not enough examinations have been administered with the new blueprint to accurately represent the question distribution (7.9 questions per category of condition seen for the 2015-2016 examinations vs 18.6 questions per category of condition seen for the 2010-2013 examinations applied in the study). Although not conclusive, a preliminary evaluation found a correlation of 0.76 between the percentages of condition questions used in this study with these percentages for the limited data available from the 2015-2016 examinations (Jonathan Vandergrift, written communication, unpublished data based on ABIM administrative data, April 11, 2017). This would suggest that the analysis presented in this study is still relevant even though some changes to the question distribution across the conditions have occurred since the blueprint enhancements were implemented. The major change that occurred in 2015 was the elimination of 12 miscellaneous/other categories, 2 of which had no questions (other ophthalmologic disorders and miscellaneous gynecology and women's health). In total, these conditions accounted for only 86 of the 3461 questions from the 2010-2013 IM-MOC examinations included in this study. Future research can both update the analysis and findings in this study and measure the effect of changes in the blueprint implemented after the 2015 blueprint review.

When considering the findings of the study, it is important to recognize that in addition to frequency of conditions seen in practice, another key dimension of content validity is the degree to which the knowledge being tested by examination questions was important in terms of its potential influence on patient care and outcomes. Because of this, in addition to the results of this study, the 2015 blueprint review also considered physician feedback regarding the importance of examination questions to practice. These data were drawn separately from a questionnaire completed by 322 practicing general internists (response rate of 13%) who rated the importance of different clinical tasks associated with subcategories of the conditions used in this study (ie, questions related to diagnosis, testing, treatment/care decisions, risk assessment/prognosis/epidemiology, and pathophysiology/basic science). Results of this questionnaire indicated that 12 (15.1% of questions) of the 72 conditions (20.5% of questions) representing 74.0% of questions found to be discordant due to too many questions had at least 1 clinical task rated as highly important. For example, pericardial disease was categorized as discordant because of a greater percentage of questions than the percentages of
Comparison of Conditions Tested by the ABIM MOC Examination vs Those Seen in Practice

Conclusions

Among questions on IM-MOC examinations from 2010-2013, 69% were concordant with conditions seen in general internal medicine practice, although some areas of discordance were identified.

The ABIM is using data in an effort to ensure that future examinations better reflect the conditions seen by practicing internists. The concordance data included in this analysis were considered by the examination committee during review of the 2015 IM-MOC examination blueprint, although the specific criteria (eg, the concordance threshold) used in this study were not. Furthermore, for subcategories of these conditions (eg, diagnosis of diabetes), the IM-MOC examination committee referenced the physician blueprint review survey described above to further adjust the examination content based on the importance and frequency of each condition reported by the physicians. Future research is needed to track whether changes in the IM-MOC examination that were informed by this and other analyses resulted in discernible improvement in the relevance of the IM-MOC examination content.

REFERENCES