Agreement Among Measures Examining Low-Value Imaging for Low Back Pain

James Henderson, PhD; Katherine Wilkinson, MS; Timothy P. Hofer, MD, MS; Robert Holleman, MPH; Mandi L. Klamerus, MPH; R. Sacha Bhatia, MD; and Eve A. Kerr, MD, MPH

ABSTRACT

OBJECTIVES: To quantify the extent of patient-level agreement among 3 published measures of low-value imaging for acute low back pain (LBP).

STUDY DESIGN: In this retrospective cohort study using commercial insurance claims from MarketScan, we assessed 3 published measures of low-value imaging for agreement in identifying LBP diagnoses (denominator), red-flag diagnoses (denominator exclusions), and imaging procedures (numerator).

METHODS: Using a cohort of patients, aged 18 to 64 years, with a diagnosis of LBP in 2014, we assessed agreement surrounding both the overuse event (imaging procedures) and inclusion in the reference population (LBP definition and exclusion diagnoses) using percent agreement and Fleiss κ among 3 overuse measures.

RESULTS: In our cohort of 1,835,620 patients with acute LBP, the 3 measures agreed 100% on the presence of acute LBP and also had excellent agreement (99%; κ = 0.98) in identifying imaging for LBP. However, there was substantial disagreement on whom to exclude for red-flag diagnoses, leading to lower agreement (75%; κ = 0.61) on whom to include in the reference population of acute LBP without red flags, among whom imaging for LBP is considered of low value.

CONCLUSIONS: Our findings demonstrate the need for further consensus surrounding how to translate guideline recommendations to administrative measures that assess overuse of imaging for acute LBP, particularly with respect to defining which patients should be excluded from the measures. This finding is also important for other overuse measures that rely on exclusions.

Am J Manag Care. 2021;27(10):438-444. doi:10.37765/ajmc.2021.88762

Low-value health care services offer limited or no benefit to patients and can even cause harm.[1] Several initiatives, including the Choosing Wisely campaign, launched in 2012 by the American Board of Internal Medicine Foundation, have focused on defining such low-value health care services[2]; several publications have attempted to assess their prevalence,[3-6] and monitoring agencies have incorporated measures of overuse in accreditation programs.[7,8] As part of this campaign, professional societies issued recommendations encouraging a period of conservative treatment before performing imaging tests for acute low back pain (LBP).[2,9] According to these recommendations, diagnostic imaging studies for LBP performed within 6 weeks of initial presentation are of low value unless there is reason to suspect that the patient's LBP is symptomatic of serious disease such as infection or cancer. This is because the vast majority of individuals with acute LBP recover within 6 weeks with conservative treatment (eg, exercise, analgesics, physical therapy) and the imaging study would not enhance treatment. Several candidate measures of this recommendation are available—including 2 from research papers examining low-value care in Medicare and a Healthcare Effectiveness Data and Information Set (HEDIS) measure—and generally expressed as a rate with a numerator counting utilization of a low-value service and a denominator counting opportunities for overuse by defining a reference population eligible for a low-value use of the service.[3-5] The reference population for these measures consists of patients with acute LBP excluding those with "red flags" for more serious disease. Thus, in defining the reference population to be included in the measures' denominators, the measures begin with a definition of acute LBP and then exclude those with concurrent or recent red-flag diagnoses. Although guidelines and recommendations establish clear clinical guidelines on what constitutes a red-flag diagnosis,[10] it is unknown to what extent existing measures agree on how to identify red-flag exclusions from administrative claims data.

It has been argued that, when constructing an overuse measure, it is preferable to take an approach that maximizes specificity in order to avoid labeling appropriate care as overuse even at the expense of sensitivity or not completely identifying all cases of overuse.[11,12]

This recommendation stems from a concern about possible unintended harms from these overuse measures if patients who may receive benefit from a service are not excluded from the measure denominator.[13] The distinction in approach is relevant as a 2014 study by Schwartz et al, which examined 26 low-value services in Medicare, found that the proportion of beneficiaries experiencing 1 or more of these low-value services fell from 42% for the more sensitive versions of their measures to 25% for the more specific versions.[4] This is particularly true if the results of the measures were to be used to restrict who is eligible to receive care or for value-based purchasing.[14,15]

TAKEAWAY POINTS

This study demonstrates the need for consensus in defining a population "eligible" for low-value imaging for acute low back pain, particularly in terms of measure exclusions, a finding likely to apply to other overuse measures relying on exclusions.

- Measures of low-value imaging for acute low back pain exhibited excellent agreement in defining overuse events (imaging procedures) and a relevant population (low back pain) but insufficient agreement in specifying diagnoses that would exclude patients with risk factors indicating potentially appropriate imaging.
- Trade-offs between sensitivity and specificity in measure definitions should be accounted for when deciding how to apply overuse measures.

Any discrepancies in measure specification resulting from differences in approach can give rise to poor agreement among measures produced by different developers. Currently, for existing administrative measures of low-value imaging in back pain, we know nothing about their agreement with respect to the specific patients identified as receiving low-value imaging studies. The impetus to decrease low-value care is strong and likely to accelerate as payers either stop reimbursing for these services or shift costs for low-value services to patients through value-based insurance designs.[16-19] Consequently, it is essential to identify and try to resolve inconsistencies in the identification of low-value care as a first step in validating these measures. We therefore examined the extent of patient-level agreement among 3 published measures of low-value imaging for acute LBP to assess the extent to which they similarly identified both LBP without red flags and the performance of low-value imaging services among this population.

METHODS

We identified and compared measures of low-value imaging for acute LBP from the literature.[3-5] For brevity, we refer to these as the Colla, HEDIS, and Schwartz measures, respectively. We selected for comparison only those measures broadly focused on imaging for acute LBP and did not consider narrower measures for a single imaging modality. We also focused on core differences in medical codes for defining the overuse event (imaging procedures) and eligible population (back pain definition and exclusion diagnoses). We first compared the measures descriptively in terms of the percentage of diagnoses or procedure codes used in common for each of the following components: LBP definition, exclusion diagnoses, and imaging procedures. We then compared the implications of these differences using a cohort of patients as described later.

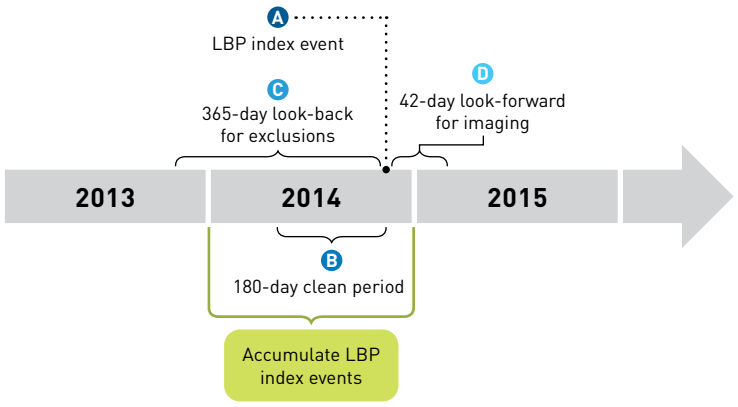
Our cohort was derived from the MarketScan Commercial Claims and Encounters Database (2013-2015; Truven Health Analytics), containing health claims from a selection of large employers, health plans, and government and public organizations and representing the medical experience of insured employees, their spouses, and dependents. Use of MarketScan data for this project was approved as not regulated by the institutional review board of the University of Michigan Medical School (HUM00141252). Using data from July 2013 through February 2015, we included all patients aged 18 to 64 years with an episode of new LBP in 2014 as defined by each of the measures, using only the first episode for each patient with multiple episodes. We then excluded enrollees without 9 months of continuous coverage (ie, month of index LBP diagnosis, 6 months prior, and 2 months after).

To focus on the agreement in the procedure codes used to identify imaging overuse and the diagnosis codes used to define a population of LBP without red flags, we standardized timing specifications that differed across the measures relating to the exposure assessment, outcome washout, and exclusion assessment time windows, preferring, where possible, a specification that was present in at least 2 of the measures. Consequently, we used a clean period (exposure assessment) of 180 days to define new acute LBP[4,5,20]; considered imaging within 42 days (6 weeks) of the index diagnosis to be of low value[3,4]; and excluded cases with a qualifying diagnosis within 365 days prior or 42 days after the index diagnosis.[3-5] (See **Figure 1** and **eAppendix Table 1** [eAppendices available at **ajmc.com**].) We used only exclusions readily identified from claims data and, privileging specificity over sensitivity,[12] required only a single physician claim for exclusions.[4,5]

For all qualifying episodes, we flagged exclusions and low-value imaging as defined by each measure, and then computed marginal rates. We also computed marginal rates by imaging modality and used logistic regression to compute odds ratios (ORs) comparing the frequency with which each measure identified low-value imaging relative to its respective denominator. We then compared the case-by-case agreement using percent agreement and Fleiss κ. Additional details on computations for the agreement statistics, including detailed summary data, are available in **eAppendix B**.

After assessing agreement among the 3 measures, the specifications of all measures were combined to form 2 "joint" measures: one maximally specific or least likely to falsely identify an LBP imaging event as being low value and the other maximally sensitive or least likely to miss a low-value LBP imaging event. The joint-specific measure uses the union of exclusions from all measures and the intersection of LBP diagnoses and LBP imaging events. In contrast, the joint-sensitive measure reverses these roles and uses

FIGURE 1. Measure Components[a]



LBP, low back pain.

[a]All 3 measures we compared share a similar structure. They all begin with an index diagnosis of new LBP (A) defined as the first LBP diagnosis within a "clean period" of 180 days (B). After identifying LBP index events, we search the prior 365 days of claim history for diagnoses indicating an event should be excluded from the denominator (C). We also look forward up to 42 days from the index event to identify claims for LBP imaging procedures (D). Although the lengths of periods B, C, and D differ among the measures, we selected a standard length for each period to emphasize comparisons between the diagnosis and procedure codes in the measure definitions.

the intersection representing exclusion diagnoses common to all measures and the union of LBP diagnoses and imaging events appearing in at least 1 measure. Estimates for these measures are constructed from summary data on the 3 primary measures. Finally, we computed projections for these joint-specific and joint-sensitive measures in the US population (aged 18-64 years with employer-sponsored insurance in 2014) using poststratification weights to adjust for differences in sex, age group, Census region, and employer relation between our cohort and this population as a whole. Weighted sums were used to project summary data on measure components for unique combinations of the 3 measures, with the projected joint-specific and joint-sensitive measures computed as above. Analyses were carried out using R versions 3.6.1 and 4.0.2 (R Foundation for Statistical Computing).

RESULTS

Comparison of Measure Components

As detailed in Figure 1, the Colla, HEDIS, and Schwartz measures evaluating overuse of imaging for acute LBP without red flags share a common structure. Each group first defines the population to which its measure applies using a new diagnosis of acute LBP as an index event, then uses diagnosis codes indicative of red-flag symptoms or history to exclude patients for whom imaging is potentially appropriate. Among this population, each measure then defines low-value imaging using procedure codes for imaging studies of the lower back within the imaging period. An LBP diagnosis is considered new if it is the first such diagnosis after a clean period without LBP diagnoses. Similarly, exclusion diagnoses are restricted to occur within a look-back period—a window of time relative to the index LBP diagnosis.

Although all 3 measures follow this structure, they differ to varying extents in the diagnosis codes used to define acute LBP and exclusions, the procedure codes defining relevant imaging, and the lengths of the clean, look-back, and imaging periods. As described in the methods and detailed in eAppendix Table 1, we standardized the lengths of these periods for the comparisons that follow, likely leading to greater agreement than comparisons without this standardization.

The disagreements in defining LBP and procedure codes for related imaging are relatively minor. In fact, Schwartz and HEDIS identically define acute LBP, whereas the Colla measure differs by 2 codes in each case. Specifically, of the 25 LBP diagnosis codes used across all measures, 23 (92%) are common to all 3 measures, 1 (4%) (spinal stenosis, lumbar region, without neurogenic claudication) is included in only the HEDIS and Schwartz measures, and 1 (4%) (Schmorl nodes lumbar region) appears in only the Colla measure. Similarly, of 22 procedure codes for imaging studies related to LBP, 20 (90.9%) are common to all 3 measures, 1 (4.5%) (MRI thoracic spine) is included in only the Schwartz measure, and 1 (4.5%) (x-ray exam C spine) appears in only the HEDIS and Schwartz measures. Of these 22 procedure codes for imaging studies, 10 (45.5%) were for plain film, 9 (40.9%) for MRI, and 3 (13.6%) for CT imaging (**eAppendix Tables 2** and **3**).

The greatest difference among measures was in the codes used to define exclusion diagnoses. Whereas 2048 (50.7%) of 4038 total codes were shared in common, 90 (2.2%) exclusion codes were shared by only 2 measures and 1900 (47.1%) were unique to a single measure. The large number of codes unique to a single measure is largely due to the inclusion of 1291 "E"-codes for "external causes of injury" in the Colla measure and 413 codes for tuberculosis in the Schwartz measure. Readers should bear in mind, however, that the number of codes is of less importance than the frequency with which those codes are applied to patients with LBP (**eAppendix Table 4**).

Marginal Rates

We found 1,835,620 patients with an episode of acute LBP in 2014 identified by at least 1 of the 3 measures and who met our coverage criteria (LBP population). This cohort was 56.8% female and fairly equally divided among age groups (**Table 1**). The percentages of cases within each measure identified as involving low-value imaging were similar for Colla and HEDIS and lower for Schwartz: 26.6% (358,992/1,350,065), 27.5% (392,625/1,425,852), and 24.0% (255,295/1,063,471), respectively.

Isolating by imaging modality and reporting low-value imaging rates for the Colla, HEDIS, and Schwartz measures, plain film accounted for most of the low-value imaging (22.8%, 23.1%, and 21.2%, respectively), followed by MRI (7.4%, 7.4%, and 4.3%), and a limited number of CT studies (0.4%, 0.4%, and 0.3%). Relative to the Schwartz measure, the Colla and HEDIS measures both identified higher rates of low-value imaging for all 3 modalities: MRI (OR [95% CI], 1.78 [1.76-1.80] and 1.78 [1.76-1.80], respectively), CT studies (OR [95% CI], 1.34 [1.29-1.41] and 1.40 [1.35-1.47]), and plain-film images (OR [95% CI], 1.09 [1.09-1.10] and 1.11 [1.11-1.12]).

TABLE 1. Demographic Characteristics for Subjects Excluded by Each Measure[a]

Group	Acute LBP, n (%)	Exclusions, n (% of LBP)		
		Colla	HEDIS	Schwartz
Female	1,045,552 (57.0)	293,458 (28.1)	240,293 (23.0)	465,252 (44.5)
Male	790,068 (43.0)	192,097 (24.3)	169,475 (21.5)	306,897 (38.8)
Aged 18-34 years	456,504 (24.9)	108,415 (23.7)	79,871 (17.5)	160,821 (35.2)
Aged 35-44 years	405,966 (22.1)	93,459 (23.0)	74,690 (18.4)	161,387 (39.8)
Aged 45-54 years	503,789 (27.4)	134,140 (26.6)	116,069 (23.0)	220,706 (43.8)
Aged 55-64 years	469,361 (25.6)	149,541 (31.9)	139,138 (29.6)	229,235 (48.8)

HEDIS, Healthcare Effectiveness Data and Information Set; LBP, low back pain.

[a]We identified 1,835,620 index events for acute LBP in 2014. The LBP column gives the distribution of these events by sex and age group. The remaining columns indicate the number and percentage of exclusions within each group for each measure.

TABLE 2. Distribution of Cases From 3 Measures for Low-Value Imaging for Acute LBP Without Red Flags[a]

Measure	LBP without red flags, n (%)	Exclusions, n (%)	Low-value imaging, n (%)
Total	1,452,231 (100)	842,426 (100)	404,611 (100)
All measures	993,194 (68.4)	383,389 (45.5)	227,733 (56.3)
Colla and HEDIS	338,250 (23.3)	7758 (0.9)	121,591 (30.1)
Colla and Schwartz	11,290 (0.8)	43,179 (5.1)	4766 (1.2)
HEDIS and Schwartz	51,229 (3.5)	7331 (0.9)	20,478 (5.1)
Colla only	7331 (0.5)	51,229 (6.1)	4902 (1.2)
HEDIS only	43,179 (3.0)	11,290 (1.3)	22,823 (5.6)
Schwartz only	7758 (0.5)	338,250 (40.2)	2318 (0.6)

HEDIS, Healthcare Effectiveness Data and Information Set; LBP, low back pain.

[a]We identified 1,835,620 index events for acute LBP in 2014. Of these, 842,426 were excluded for red flags by 1 or more measures. The distribution of these exclusions across combinations of measures is presented under "exclusions." The remaining 993,194 cases are included in the denominator of all measures. In total, 1,452,231 cases appear in the denominator of at least 1 measure, with the distribution of these cases shown under "LBP without red flags." Among these, 404,611 were flagged for "low-value imaging" and appear in at least 1 numerator. Differences in exclusions rather than how imaging for LBP is defined account for the majority of the differences in cases flagged for low-value imaging (data not shown).

Index Events

Using a standardized clean period, there was near-perfect agreement among measures in identifying acute LBP, with 100% of patients identified by all 3 measures. We say "near" perfect because although the measures agree that all 1,835,620 had an LBP episode with index diagnosis in 2014, there were a small number (4088) of disagreements as to the precise *date* of the index diagnosis within 2014, owing to differences in LBP diagnoses described earlier.

Imaging for Acute LBP

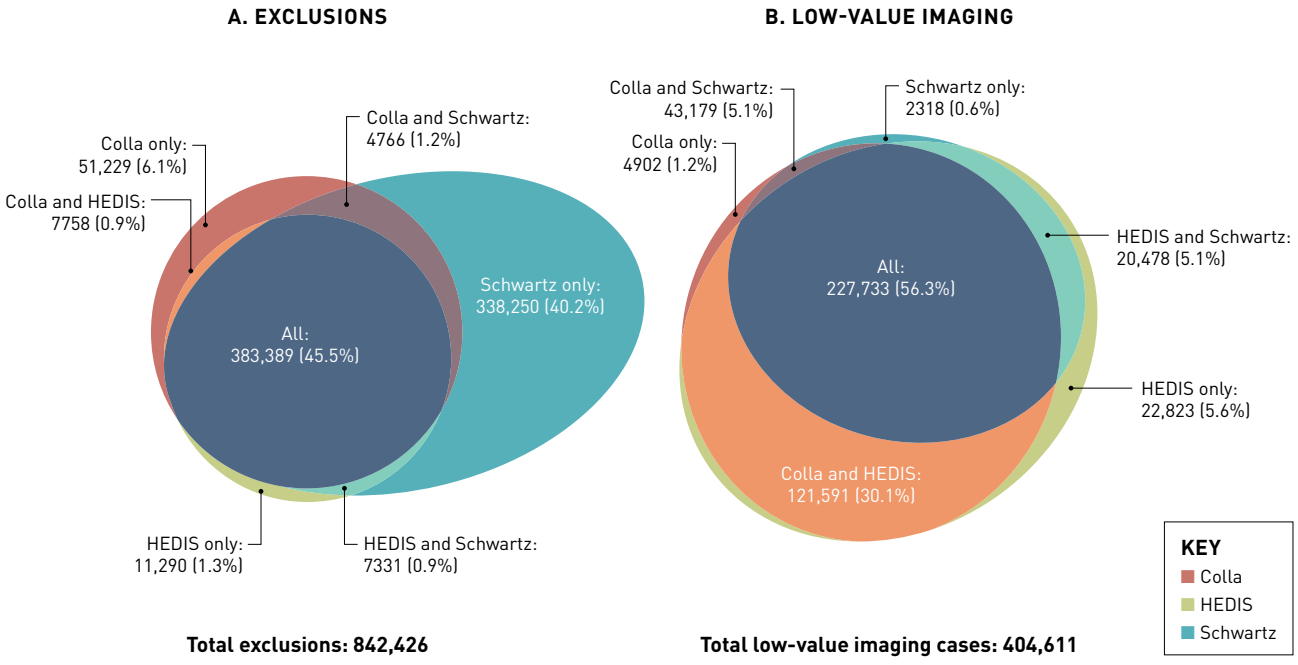
There was also excellent agreement (99.0%; $\kappa = 0.98$) on what constitutes imaging for LBP. Among the LBP population, 70.7% (1,298,528/1,835,620) had no imaging claims and 28.2% (518,158/1,835,620) had claims for imaging procedures included in all 3 measures. Only 1.0% (18,934/1,835,620) had imaging claims for procedures on which the measures disagree, which were primarily related to small differences in Current Procedural Terminology codes used to capture imaging procedures as described earlier. Moreover, the 518,158 patients with consensus imaging represent 96.5% of the 537,092 patients with imaging claims on any of the 3 measures. Among the latter group, 3.5% (18,536/537,092) had imaging claims included in only the HEDIS and Schwartz measures but not the Colla measure, whereas other combinations accounted for less than 0.1% (398/537,092).

Exclusions for Red-Flag Diagnoses and LBP Without Red Flags Population

However, there was substantial disagreement as to which red-flag diagnoses were used when excluding from the denominator patients for whom imaging is potentially appropriate. Consequently, there was lower agreement (75.8%; $\kappa = 0.62$) on which episodes were included in the denominator. Although all 3 measures agreed on the absence of exclusions in 54.1% (993,194/1,835,620) of cases, only 20.9% (383,389/1,835,620) of cases were excluded by all 3 measures, representing only 45.5% (383,389/842,426) of those excluded by at least 1 measure (see **Table 2** and **Figure 2 [A]**).

The Schwartz measure had by far the largest number of exclusions. Among patients excluded by at least 1 measure (842,426), the Schwartz measure excluded 91.7% (772,149) and *uniquely* excluded 40.2% (338,250). The Colla (51,229) and HEDIS (11,290) measures had fewer unique exclusions. The most frequently encountered exclusion diagnoses about which the measures disagree are listed in **Table 3**. Notably, the red-flag diagnoses contributing most to the larger number of exclusions from the Schwartz measure relative to the others are malaise and fatigue (780.79), unspecified thoracic or lumbosacral neuritis or radiculitis (724.4), unspecified anemia (285.9), and unspecified fever (780.60). Examining the measures in pairs, there is greater agreement on whom to include in the denominator between the Colla and HEDIS measures (93.8%; $\kappa = 0.83$) than between Colla and Schwartz (78.0%; $\kappa = 0.51$) or HEDIS

FIGURE 2. Venn Diagrams Illustrating Distribution of Exclusions and Low-Value Imaging[a]



HEDIS, Healthcare Effectiveness Data and Information Set; LBP, low back pain.

[a]The diagram on the left (A) compares measures in terms of which patients with LBP each measure excludes from the denominator representing acute LBP without red flags. A total of 45.9% (842,426/1,835,620) of LBP events are excluded by at least 1 measure, but only 45.5% (383,389/842,426) of these are excluded by all 3 measures. Refer to Table 3 for a summary of the diagnoses differentiating these exclusion groups. In contrast, the diagram on the right (B) compares measures in terms of instances of low-value imaging for LBP. Recall that low-value imaging is any imaging for LBP done among those *not* excluded by a measure. A total of 22.0% (404,611/1,835,620) of cases were flagged for low-value imaging by at least 1 measure, with all 3 measures agreeing that low-value imaging took place for 56.3% (227,733/404,611) of these cases.

and Schwartz (78.2%; $\kappa = 0.50$) because of the higher frequency of exclusions in the Schwartz measure.

Low-Value Imaging for LBP

Agreement on which patients received low-value imaging for acute LBP (ie, those who received imaging and were not excluded from the denominators of the measures due to red-flag diagnoses) was 90.4% ($\kappa = 0.79$). Consensus agreement that a patient received low-value care represents only 56.3% (227,733/404,611) of the cases identified as low value by at least 1 measure. Among the remainder, 30.1% (121,591/404,611) were identified as low value by the Colla and HEDIS measures, but not Schwartz, reflecting the larger number of exclusions identified by Schwartz as described above. Relative to the entire LBP population, 78.0% (227,733/1,835,620) had no low-value imaging and 12.4% (227,733/1,835,620) of patients had claims for imaging procedures identified as low value by all 3 measures. Examining pairwise agreement on which patients received low-value imaging, there is again greater agreement between the Colla and HEDIS measures (97.1%; $\kappa = 0.91$) than between Colla and Schwartz (91.9%; $\kappa = 0.71$) or HEDIS and Schwartz (91.7%; $\kappa = 0.72$) (Table 2 and **Figure 2 [B]**). Additionally, restricting attention to subsets of patients who received imaging of a specific modality, as identified by at least 1 measure, there was greater agreement as to when an x-ray (70.1%; $\kappa = 0.56$) or CT scan (69.3%; $\kappa = 0.57$) was low value than for MRI (53.8%; $\kappa = 0.38$).

Joint Measures

Finally, to better understand the specificity-sensitivity tradeoffs embodied in the Colla, HEDIS, and Schwartz measures, we combined specifications from the 3 measures to form either more sensitive or more specific "joint" measures. The joint-specific measure maximizes specificity by using the union of exclusion diagnoses from all 3 measures, the intersection of LBP diagnoses, and the intersection of imaging procedures. In contrast, we define the joint-sensitive measure to maximize sensitivity by using the intersection of exclusion diagnoses, the union of LBP diagnoses, and the union of imaging procedures.

As index diagnoses for identifying LBP have near-perfect agreement across the 3 measures, the joint-specific and joint-sensitive measures can be closely approximated using figures from Table 2.

The joint-specific measure identifies 227,733 cases of low-value imaging from 993,194 patients with LBP and no red flags, resulting in a low-value testing rate of 22.9%. In contrast, the joint-sensitive measure identifies 404,611 instances of low-value imaging from 1,452,231 qualifying patients, resulting in a rate of 27.9%. Although the 5% difference in rates does not seem particularly consequential, projecting the volume of patients receiving low-value imaging from the analyzed cohort to the population represented by our data (those aged 18-64 years with employer-sponsored insurance), the difference between the maximally sensitive and the maximally specific measures represents a difference of 580,010 (751,416 vs 1,331,426) more patients receiving low-value imaging.

TABLE 3. Most Frequent Exclusion Codes Not Common to All 3 Measures[a]

ICD-9 exclusion	Measures using exclusion	Patients excluded, n (%)
780.79: Other malaise and fatigue	Schwartz	247,643 (29.4%)
724.4: Thoracic or lumbosacral neuritis or radiculitis, unspecified	Schwartz	177,901 (21.1%)
285.9: Anemia, unspecified	Schwartz	74,462 (8.8%)
780.60: Fever, unspecified	Schwartz	53,473 (6.3%)
729.2: Neuralgia, neuritis, and radiculitis, unspecified	Colla, Schwartz	33,902 (4.0%)
V108.3: Personal history of other malignant neoplasm of skin	HEDIS	18,412 (2.2%)
795.01: Papanicolaou smear of cervix with atypical squamous cells of undetermined significance	Colla	17,585 (2.1%)
783.21: Loss of weight	Schwartz	16,758 (2.0%)
722.93: Other and unspecified disc disorder, lumbar region	HEDIS	16,490 (2.0%)
V103: Personal history of malignant neoplasm of breast	Colla, HEDIS	14,999 (1.8%)
795.03: Papanicolaou smear of cervix with low grade squamous intraepithelial lesion	Colla	9889 (1.0%)
E8889: Unspecified fall	Colla	9067 (1.1%)

HEDIS, Healthcare Effectiveness Data and Information Set; *ICD-9*, *International Classification of Diseases, Ninth Revision*.

[a]Shown here are those diagnoses appearing in the history of at least 1% of the 842,426 patients excluded by at least 1 measure, excluding diagnoses for which all 3 measures agree. The numbers here reflect the number of unique patients with each exclusion in their history. Many patients counted here are not uniquely excluded by the listed measure(s) as they may also have other exclusions in their histories. For instance, whereas 247,643 patients have a diagnosis of 780.79 (unique to Schwartz) during the look-back period, only 167,012 of these patients are excluded solely from the Schwartz measure.

DISCUSSION

To the best of our knowledge, this is the first peer-reviewed study to assess patient-level agreement among measures of low-value care for a common overuse event. We found that all 3 measures provide comparable marginal rates of low-value imaging for LBP among those included in a given measure. However, owing to differences on which red-flag diagnoses to use for excluding patients, there was substantial disagreement about who has received low-value imaging. A total of 176,878 cases of imaging for LBP were considered low value by 1 or 2 measures but not all measures. This represents 9.6% of the LBP population and 43.7% of cases considered low value by 1 or 2 but not all measures.

Overuse measures have the potential to be applied in different ways. A health system could use such measures to track at a high level the proportion of patients who may be receiving unnecessary services, in order to better tailor quality improvement initiatives. In this case, a sensitive measure may be appropriate. On the other hand, if the measure were to be applied prospectively in clinical decision support tools, as promoted by CMS,[21,22] it may be more appropriate to use specific measures to ensure that patients who need imaging are not prevented from receiving it. Exclusion distinctions are also critical to consider when measures are incorporated into value-based payments,[14-19] as some health systems with more complex patients (eg, with potential for more exclusions) could be penalized if more sensitive measures to track overuse were implemented. More concerning, incentivizing performance using highly sensitive measures could lead to underuse of needed services. Differences in the use of exclusions of the magnitude that we report reflect fundamental disagreement about what represents low-value care and need to be considered when deciding how to apply overuse measures.

Limitations

While recognizing that any claims-based measure has limitations, we believe this work demonstrates the need for further consensus surrounding high-value exclusions for imaging of LBP and other measures of low-value care. It is further noteworthy that use of administrative data to capture these exclusions has significant limitations and that being able to include more relevant clinical information from the electronic health record may obviate the need for use of general codes (eg, unspecified fever) to capture exclusions.

It is noteworthy that the measures we compared on a cohort of commercially insured patients aged 18 to 64 years were developed for and/or against different age groups: the Colla and Schwartz measures for Medicare patients 65 years and older and the HEDIS measure for patients aged 18 to 50 years. Despite this difference, the comparisons presented remain valid because common reasons justifying immediate imaging and thus measure exclusion—major neurological deficits or signs/risk factors for cancer, spinal infection, or cauda equina syndrome—are relevant for patients of all ages.[16] Indeed, according to a guideline from the American College of Physicians, age is primarily relevant as a risk factor justifying imaging only after a period of conservative treatment; such delayed imaging is not covered by the measures that we compared. Therefore, the differences in the exclusions do not solely reflect the age of the targeted populations. However, it is important to acknowledge that the strength of evidence needed to justify immediate imaging depends on the baseline prevalence of the suspected condition.

Therefore, measures designed for those older than 65 years may define more permissive exclusions for conditions, such as cancer, for which age is also a risk factor. For example, in the guideline referenced above, "age > 50" and "unexplained weight loss" are each considered "weaker risk factors for cancer" justifying, individually, only delayed imaging for LBP. Taken together, these 2 factors could potentially justify immediate imaging.

CONCLUSIONS

For the next generation of overuse measures to be effective, health care providers must view them as valid and meaningful.[23] It is therefore necessary that overuse measures specify numerators and denominators that are clearly defined, include all clinically appropriate exclusions, and are subject to the broadest possible consensus—particularly in relation to exclusions. This will help ensure that appropriate use of medical services is not misclassified as inappropriate use so that overuse measures help, and do not hinder, efforts to improve value in health care. ■

Author Affiliations: Consulting for Statistics, Computing, and Analytics Research, University of Michigan (JH), Ann Arbor, MI; VA Center for Clinical Management Research (JH, KW, TPH, RH, MLK, EAK), Ann Arbor, MI; Institute for Healthcare Policy and Innovation and Department of Internal Medicine, University of Michigan (TPH, EAK), Ann Arbor, MI; Institute for Health Systems Solutions and Virtual Care, Women's College Hospital (RSB), Toronto, Ontario, Canada; Institute of Health Policy, Management and Evaluation, University of Toronto (RSB), Toronto, Ontario, Canada.

Source of Funding: Funding for this work was provided by VA Health Services Research & Development (USA 18-175). The funders had no role in the design and conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Author Disclosures: The authors report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (JH, EAK); acquisition of data (JH, RH, EAK); analysis and interpretation of data (JH, KW, TPH, RH, RSB, EAK); drafting of the manuscript (JH, TPH, MLK, RSB); critical revision of the manuscript for important intellectual content (JH, KW, TPH, MLK, RSB, EAK); statistical analysis (JH, KW, TPH, RH); and administrative, technical, or logistic support (MLK, EAK).

Address Correspondence to: Eve A. Kerr, MD, MPH, University of Michigan, 2800 Plymouth Rd, Bldg 16, Ann Arbor, MI 48109-2800. Email: ekerr@med.umich.edu.

REFERENCES

1. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. The National Academies Press; 2001.
2. Cassel CK, Guest JA. Choosing wisely: helping physicians and patients make smart decisions about their care. *JAMA*. 2012;307(17):1801-1802. doi:10.1001/jama.2012.476
3. Colla CH, Morden NE, Sequist TD, Schpero WL, Rosenthal MB. Choosing wisely: prevalence and correlates of low-value health care services in the United States. *J Gen Intern Med*. 2015;30(2):221-228. doi:10.1007/s11606-014-3070-z
4. Schwartz AL, Landon BE, Elshaug AG, Chernew ME, McWilliams JM. Measuring low-value care in Medicare. *JAMA Intern Med*. 2014;174(7):1067-1076. doi:10.1001/jamainternmed.2014.1541
5. 2017 Quality Rating System Measure Technical Specifications. CMS. September 2016. Accessed March 26, 2019. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/2017_QRS-Measure_Technical_Specifications.pdf
6. Segal JB, Bridges JFP, Chang HY, et al. Identifying possible indicators of systematic overuse of health care procedures with claims data. *Med Care*. 2014;52(2):157-163. doi:10.1097/MLR.0000000000000052
7. Chassin MR, Loeb JM, Schmaltz SP, Wachter RM. Accountability measures—using measurement to promote quality improvement. *N Engl J Med*. 2010;363(7):683-688. doi:10.1056/NEJMsb1002320
8. Imaging efficiency measures. QualityNet. Accessed April 30, 2019. https://qualitynet.cms.gov/outpatient/measures/imaging-efficiency
9. Imaging tests for lower-back pain. Choosing Wisely. 2017. Accessed March 26, 2019. http://www.choosingwisely.org/patient-resources/imaging-tests-for-back-pain/
10. Chou R, Qaseem A, Owens DK, Shekelle P; Clinical Guidelines Committee of the American College of Physicians. Diagnostic imaging for low back pain: advice for high-value health care from the American College of Physicians. *Ann Intern Med*. 2011;154(3):181-189. doi:10.7326/0003-4819-154-3-201102010-00008
11. Baker DW, Qaseem A, Reynolds PP, Gardner LA, Schneider EC; American College of Physicians Performance Measurement Committee. Design and use of performance measures to decrease low-value services and achieve cost-conscious care. *Ann Intern Med*. 2013;158(1):55-59. doi:10.7326/0003-4819-158-1-201301010-00560
12. Saini SD, Powell AA, Dominitz JA, et al. Developing and testing an electronic measure of screening colonoscopy overuse in a large integrated healthcare system. *J Gen Intern Med*. 2016;31(suppl 1):53-60. doi:10.1007/s11606-015-3569-y
13. Mathias JS, Baker DW. Developing quality measures to address overuse. *JAMA*. 2013;309(18):1897-1898. doi:10.1001/jama.2013.3588
14. Fendrick AM, Chernew ME. Value-based insurance design: aligning incentives to bridge the divide between quality improvement and cost containment. *Am J Manag Care*. 2006;12(Spec No. 12):SP5-SP10.
15. Gibson TB, Maclean RJ, Chernew ME, Fendrick AM, Baigel C. Value-based insurance design: benefits beyond cost and utilization. *Am J Manag Care*. 2015;21(1):32-35.
16. Keats JP. Curtailing utilization of low-value medical care. *Am J Accountable Care*. 2019;7(2):24-25.
17. Gruber J, Maclean JC, Wright B, Wilkinson E, Volpp KG. The effect of increased cost-sharing on low-value service use. *Health Econ*. 2020;29(10):1180-1201. doi:10.1002/hec.4127
18. Barthold D, Basu A. A scalpel instead of a sledgehammer: the potential of value-based deductible exemptions in high-deductible health plans. *Health Affairs*. June 18, 2020. Accessed September 11, 2020. https://www.healthaffairs.org/do/10.1377/hblog20200615.238552/full/
19. Dhruva SS, Redberg RF. A successful but underused strategy for reducing low-value care: stop paying for it. *JAMA Intern Med*. 2020;180(4):532. doi:10.1001/jamainternmed.2019.7142
20. Pham HH, Landon BE, Reschovsky JD, Wu B, Schrag D. Rapidity and modality of imaging for acute low back pain in elderly patients. *Arch Int Med*. 2009;169(10):972-981. doi:10.1001/archinternmed.2009.78
21. Timbie JW, Hussey PS, Burgette LF, et al. Medicare imaging demonstration final evaluation: report to Congress. *Rand Health Q*. 2015;5(1):4.
22. Appropriate Use Criteria Program. CMS. Accessed March 26, 2019. https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/appropriate-use-criteria-program/index.html
23. MacLean CH, Kerr EA, Qaseem A. Time out—charting a path for improving performance measurement. *N Engl J Med*. 2018;378(19):1757-1761. doi:10.1056/NEJMp1802595

Visit **ajmc.com/link/88762** to download PDF and eAppendix

eAppendix A

eAppendix Table 1. *Determination of standardized measure components.* We standardized specification of measure components other than the medical diagnosis codes used to determine measure eligibility and the procedure codes used to capture the overuse event.

Measure Component	Colla	HEDIS	Schwartz	Standardized
Clean period	Study period	6 months	6 months[3]	6 months
Exclusion period	12 months[1]	12 months[2]	Study period[4]	12 months
Imaging Period	42 days (6 weeks)	28 days	42 days (6 weeks)	42 days

Notes: 1. Colla et al use all available history for cancer exclusions and require two physician claims for exclusions other than trauma. 2. The HEDIS measure specifies the use of all available history for exclusions related to cancer or human immunodeficiency virus (HIV), and limits exclusions for trauma to three months. 3. We inferred the use of a 6-month clean period by Schwartz et al from their reference to (Pham, 2009). 4. The length of the exclusion period in Schwartz et al was not explicitly stated, but we inferred it to be all history during the study period.

eAppendix Table 2. *Comparison of low-back pain diagnoses.* For each ICD9 diagnosis code range (columns) this table shows the number of ICD9 codes shared by a given measure combination (rows).

	ICD9 Diagnoses for Low Back Pain (n)			
Measures	**721-722**	**724**	**738-739**	**846-847**
All measures	6	7	3	7
HEDIS & Schwartz	-	1	-	-
Colla only	1	-	-	-

eAppendix Table 3. *Comparison of imaging procedures.* For each imaging modality (columns) this table shows the number of procedure codes shared by a given measure combination (rows).

	Imaging Modality for LBP Imaging Studies		
Measures	Plain Film (XR)	Magnetic Resonance (MR)	Computed Tomography (CT)
All measures	9	8	3
HEDIS & Schwartz	1	-	-
Schwartz only	-	1	-

eAppendix Table 4. *Comparison of red-flag exclusion diagnoses.* For each category of red-flag exclusions (rows) this table shows the number of ICD9 codes shared by a given measure combination (columns). C = Colla, H = HEDIS, S = Schwartz.

	Measures						
Category	**All**	**C&H**	**C&S**	**H&S**	**C**	**H**	**S**
Cancer	349	77	-	-	-	57	-
Malignant neoplasms	452	-	-	-	-	28	-
Secondary neuroendocrine tumor	-	1	-	-	-	-	-
General signs and symptoms, unspecified anemia	-	-	-	-	-	-	12
Trauma, External causes of injuries	1214	-	-	-	1291	1	-
Myelopathy, Neuritis, Radiculitis, Radiculopathy	-	-	-	-	-	-	5
Neurological impairment	1	-	1	-	-	-	-
Kidney transplant	-	-	-	-	-	1	-
HIV, immune deficiencies	-	1	-	-	3	1	-
Tuberculosis	-	-	-	7	-	-	413
Spinal infection	-	2	-	-	-	1	-
Osteomyelitis	-	-	-	1	-	-	69
Endocarditis	-	-	-	-	-	-	3
Septicemia	-	-	-	-	-	-	15
IV Drug Abuse	32	-	-	-	-	-	-

Supplementary Data and Methods

Summary data and detailed methods for computing the agreement statistics are in **eAppendix B** (https://jbhender.github.io/research/hsr/lbp_agree/eAppendixB.html).

Code

Supporting material for “Agreement measures examining low-value imaging for low back pain”

James Henderson, Katherine Wilkinson, Timothy P. Hofer, Rob Holleman, Mandi L. Klamerus, R. Sacha Bhatia, Eve A. Kerr

October 02, 2020

Overview

In this appendix, we provide counts for all measure components for each of the eight possible combinations of the three measures considered. These counts are sufficient to reproduce all statistics and figures with the exception of: (1) the demographics in table 1, (2) the specific exclusions listed in table 2, and (3) the population projections in “Joint Measures”.

This file also contains the code used to compute agreement statistics from the counts provided. To view this code, use the `Code` buttons along the right side of the page or choose “Show All Code” on the upper right. The source code for the html file you are reading is available as an executable R script, `Appendix2.R`.

Tables

```
# 79: ------------------------------------------------------------------------
```

```
# libraries: ------------------------------------------------------------------
suppressPackageStartupMessages({
  library(tidyverse); library(data.table); library(lubridate)
})

# aggregated data: ------------------------------------------------------------
path = '~/github/USvCA/data/lbp/'

for ( ff in c("idx_tab",  "den_tab",  "num_tab",  "im_tab") ) {
   file = sprintf('%s/%s.csv', path, ff)
   data = fread(file)
   assign(ff, data)
}
rm(data, file)
```

```
# map count of TRUE to number of agreements: ---------------------------------
map_k = function(k) {
  if ( k %in% c(0, 3) ) return(3)
  if ( k %in% c(1, 2) ) return(1)
}
mapk = function(k) sapply(k, map_k)

# Agreement: ------------------------------------------------------------------

## Global parameters
N = sum(idx_tab$N)
n = 3
k = 2

## Denominator table

### Kappa
den_tab[, w := mapk(S_ep + C_ep + H_ep)]
Pbar =  with(den_tab, sum(w * N) / {3 * sum(N)} )
p1 = with(den_tab, sum( {S_ep + C_ep + H_ep} * N ) / {sum(N) * 3})
Pe = p1^2 + {1 - p1}^2
kappa_den = {Pbar - Pe} / {1 - Pe}

### Percent Agreement
p = den_tab[, sum(N[{S_ep & C_ep & H_ep} |  {!S_ep & !C_ep & !H_ep}]) / sum(N)]

### Summary statistics
den_stats = list(p = p, kappa = kappa_den, p1 = p1, Pe = Pe, Pbar = Pbar)
```

```
mapl = function(x){
  ifelse(x, 'Yes', 'No')
}
excl_cap = "**Table A1.** *Exclusion counts by measure combination.*"

den_tab[order(C_ep, H_ep, S_ep),
        .(Colla = mapl(!C_ep),
          HEDIS = mapl(!H_ep),
          Scwhartz = mapl(!S_ep),
          N = format(N, big.mark = ','),
          `# Agreements` = w
        )] %>%
  knitr::kable(format = 'html', caption = excl_cap) %>%
  kableExtra::kable_styling("striped", full_width = TRUE) %>%
  kableExtra::add_header_above(c("Excluded by Measure" = 3, " " = 1, " " = 1))
```

Table A1. *Exclusion counts by measure combination.*

Excluded by Measure				
Colla	**HEDIS**	**Scwhartz**	**N**	**# Agreements**
Yes	Yes	Yes	383,389	3
Yes	Yes	No	7,758	1
Yes	No	Yes	43,179	1
Yes	No	No	51,229	1
No	Yes	Yes	7,331	1
No	Yes	No	11,290	1
No	No	Yes	338,250	1
No	No	No	993,194	3

```
# Imaging: ----------------------------------------------------------------
```

```
im_tab[, w := mapk(Schwartz_im + Colla_im + Hedis_im)]
Pbar = with(im_tab, sum(w * N) / {3 * sum(N)})
p1 = with(im_tab, sum( {Schwartz_im + Colla_im + Hedis_im} * N) / {sum(N) * 3})
Pe = p1^2 + {1 - p1}^2
kappa_im = {Pbar - Pe} / {1 - Pe}

p = im_tab[ , sum(N[{Schwartz_im & Colla_im & Hedis_im} |
{!Schwartz_im & !Colla_im & !Hedis_im}]) / sum(N)]
im_stats = list(p = p, kappa = kappa_im, p1 = p1, Pe = Pe, Pbar = Pbar)
```

```
im_cap = "**Table A2.** *Imaging counts by measure combination.*"
```

```
im_tab[order(Colla_im, Hedis_im, Schwartz_im),
       .(Colla = mapl(Colla_im),
         HEDIS = mapl(Hedis_im),
         Scwhartz = mapl(Schwartz_im),
         N = format(N, big.mark = ','),
         `# Agreements` = w
       )] %>%
 knitr::kable(format = 'html', caption = im_cap) %>%
 kableExtra::kable_styling("striped", full_width = TRUE) %>%
```

```
kableExtra::add_header_above(c("LBP Imaging by Measure" = 3, " "= 1, " "= 1))
```

Table A2. *Imaging counts by measure combination.*

LBP Imaging by Measure				
Colla	**HEDIS**	**Scwhartz**	**N**	**# Agreements**
No	No	No	1,298,528	3
No	No	Yes	1	1
No	Yes	No	33	1
No	Yes	Yes	18,536	1
Yes	No	No	353	1
Yes	No	Yes	11	1
Yes	Yes	Yes	518,158	3

```
# Numerator: --------------------------------------------------------------------
```

```
num_tab[, w := mapk(S_im + C_im + H_im)]
Pbar = with(num_tab, sum(w * N) / {3 * sum(N)})
p1 = with(num_tab, sum( {S_im + C_im + H_im} * N) / {sum(N) * 3})
Pe = p1^2 + {1 - p1}^2
kappa_num = {Pbar - Pe} / {1 - Pe}

p = num_tab[, sum(N[{S_im & C_im & H_im} |  {!S_im & !C_im & !H_im}]) / sum(N)]
num_stats = list(p = p,  kappa = kappa_num, p1 = p1, Pe = Pe, Pbar = Pbar)
```

```
n0 = "**Table A3.** *Numerator counts by measure combination.* "
```

```
n1 = "A patient is in the numerator if both recieved imaging and are included "
n2 = "in the denominator, i.e. not excluded."
num_cap = paste0(n0, n1, n2)

num_tab[order(C_im, H_im, S_im),
       .(Colla = mapl(C_im),
         HEDIS = mapl(H_im),
         Scwhartz = mapl(S_im),
         N = format(N, big.mark = ','),
         `# Agreements` = w
       )] %>%
```

```
knitr::kable(format = 'html', caption = num_cap) %>%
kableExtra::kable_styling("striped", full_width = TRUE) %>%
kableExtra::add_header_above(
  c("Low-value LBP Imaging by Measure" = 3, " " = 1, " " = 1)
)
```

Table A3. *Numerator counts by measure combination.* A patient is in the numerator if both recieved imaging and are included in the denominator, i.e. not excluded.

Low-value LBP Imaging by Measure				
Colla	**HEDIS**	**Scwhartz**	**N**	**# Agreements**
No	No	No	1,431,009	3
No	No	Yes	2,318	1
No	Yes	No	22,823	1
No	Yes	Yes	20,478	1
Yes	No	No	4,902	1
Yes	No	Yes	4,766	1
Yes	Yes	No	121,591	1
Yes	Yes	Yes	227,733	3

Additional Methods

Computation of percent agreement

Percent agreement is computed as,

$$ P_A = \frac{N_A}{N} \times 100 $$

where N is the total number of cases and N_A is the number of cases for which all three measures agree (all “Yes” or all “No”) that a given definition has, or has not, been met for: an index LBP diagnosis (cohort entry event), an exclusion diagnosis indicating potentially appropriate LBP imaging (exclusions), or an LBP imaging event (outcome).

Computation of κ

Fleiss's κ is an agreement statistic that reports the relative amount by which the observed agreement between pairs of measures exceeds that expected if all measures assigned values uniformly at random according to the aggregrate marginal probabilities. Specifically, if Y_{nk} is a binary variable indicating membership for case n $((n = 1, \cdots, N))$ in a particular component as defined by measure k $((k = 1, \cdots, K))$, then

$$ \kappa = \frac{P_O - P_E}{1 - P_E}. $$

In the definition above, the observed agreement probability, P_O, is computed as,

$$ P_O = \frac{1}{N \times \left(K, 2\right)} \sum_{n=1}^N \sum_{k \ne j} Y_{nk} = Y_{nj} $$

where, $((K, 2))$ is the number of combinations of K measures taken 2 at a time; for $K = 3$ as here, we have $((3, 2) = 3)$. P_O can also be computed directly from tables A1-A3 by multiplying "N" by "# Agreements" and then dividing by the maximum possible number of agreements, 3 $\times$ 1,835,620 = 5,506,860.

Similarly, the expected agreement, P_E, is computed as,

$$ P_E = p^2 + (1-p)^2, \quad p = \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K Y_{nk}. $$

Table A4 provides these statistics for LBP measure exclusions, LBP measure imaging events, and LBP measure numerators.

```
p_cap = "**Table A4.** Components of agreement statistics."
```

```
cbind( data.table("Component" = c("Exclusions", "Imaging", "Numerator")),
 rbind(as.data.table(den_stats),
     as.data.table(im_stats),
     as.data.table(num_stats)
 ))[,.(Component, `$P_A$` = p, `$P_O$`=Pbar, `$p$` = p1,
       `$P_E$` = Pe, `$\\kappa$` = kappa)
 ] %>%
 knitr::kable(format='html', digits = 2, caption = p_cap) %>%
 kableExtra::kable_styling("striped", full_width = TRUE)
```

Table A4. Components of agreement statistics.

Component	**P_A**	**P_O**	**p**	**P_E**	**κ**
Exclusions	0.75	0.83	0.70	0.58	0.61
Imaging	0.99	0.99	0.29	0.59	0.98
Numerator	0.90	0.94	0.18	0.70	0.79