OPEN

Validation of Diagnostic Groups Based on Health Care Utilization Data Should Adjust for Sampling Strategy

Geneviève Cadieux, MD, PhD,† Robyn Tamblyn, PhD,†‡ David L. Buckeridge, MD, PhD,†§ and Nandini Dendukuri, PhD†*

Objective: Valid measurement of outcomes such as disease prevalence using health care utilization data is fundamental to the implementation of a "learning health system." Definitions of such outcomes can be complex, based on multiple diagnostic codes. The literature on validating such data demonstrates a lack of awareness of the need for a stratified sampling design and corresponding statistical methods. We propose a method for validating the measurement of diagnostic groups that have: (1) different prevalences of diagnostic codes within the group; and (2) low prevalence.

Methods: We describe an estimation method whereby: (1) low-prevalence diagnostic codes are oversampled, and the positive predictive value (PPV) of the diagnostic group is estimated as a weighted average of the PPV of each diagnostic code; and (2) claims that fall within a low-prevalence diagnostic group are oversampled relative to claims that are not, and bias-adjusted estimators of sensitivity and specificity are generated.

Application: We illustrate our proposed method using an example from population health surveillance in which diagnostic groups are applied to physician claims to identify cases of acute respiratory illness.

Conclusions: Failure to account for the prevalence of each diagnostic code within a diagnostic group leads to the underestimation of the PPV, because low-prevalence diagnostic codes are more likely to be false positives. Failure to adjust for oversampling of claims that fall within the low-prevalence diagnostic group relative to those that do not leads to the overestimation of sensitivity and underestimation of specificity.

From the *Dalla Lana School of Public Health, University of Toronto, Toronto, ON; †Department of Epidemiology, Biostatistics and Occupational Health, McGill University; ‡Direction de la Santé Publique de Montréal; and §Department of Medicine, McGill University, Montreal, QC, Canada.

Supported by the Canadian Institutes for Health Research and the McGill University Health Centre Research Institute.

The authors declare no conflict of interest.

Reprints: Geneviève Cadieux, MD, PhD, Dalla Lana School of Public Health, University of Toronto, 155 College St., 6th Floor, Toronto, ON, Canada M5T 3M7. E-mail: genevieve.cadieux@mail.utoronto.ca.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website, www.lww-medicalcare.com.

ISSN: 0025-7079/17/5508-0e59

Key Words: verification bias, validation studies, diagnostic groups, healthcare utilization data, stratified sampling, surveillance

(*Med Care* 2017;55: e59–e67)

Recent interest toward developing a "learning health system," in which new knowledge is generated as an integral byproduct of health care delivery,[1] has prompted calls for improved capacity to draw timely inference from health care data.[2–4] Major investments in health information technology, including the development and widespread implementation of electronic health records,[5] have created new streams of health data and added richness and complexity to existing data streams. As this wealth of new data becomes available, methods to infer from these data important metrics, such as disease prevalence, must be validated before they are used to guide decisions within a "learning health system."

In addition to playing a critical role in achieving the vision of a "learning health system," the ability to draw sound inference from comprehensive and high-quality health care utilization data (ie, data generated as a corollary of health service delivery) is a cornerstone of quality improvement, pharmacosurveillance, public health practice and policymaking, and health-related research in a variety of disciplines. Diagnostic codes are among the most widely used data elements in health care utilization data; they are used to identify patient populations of interest,[6–8] to assess the presence of risk factors,[9,10] to adjust for comorbidity[11–14] and case-mix,[14–16] to monitor health outcomes in the population,[17,18] and even to inform individual patient care at the time of the clinical encounter.[19] However, because diagnostic codes in health care utilization data are generated as a byproduct of health services delivery (eg, to enable fee-for-service remuneration), before using them for another purpose, it is essential to determine whether these data are sufficiently sensitive and specific for that purpose. Typically, when these data are used to measure something about a particular condition, a group of diagnostic codes associated with the condition is defined and validated. Diagnostic groups are generally used because individual diagnostic codes are too "fine-grained" for most purposes.

The scientific literature is replete with validation studies of diagnostic groups in health care utilization data. However, the quality of their methodology is highly variable, ranging from cross-correlations between 2 time-series[20,21] to record-level comparison with other data sources including

TABLE 1. Example of a Diagnostic Group Made up of Diagnostic Codes, Each With a Different Population Prevalence: The CDC's Diagnostic Group for "Respiratory Syndrome"

Diagnostic group	Conceptual definition for the diagnostic group	Examples of corresponding ICD-9-CM diagnostic codes	Number of visits with this ICD-9-CM code[1]	Prevalence of this ICD-9-CM code[2]
Respiratory syndrome (35)	■ ACUTE infection of the upper and/ or lower respiratory tract (from the oropharynx to the lungs, includes otitis media). ■ SPECIFIC diagnosis of acute respiratory tract infection (RTI) such as pneumonia due to parainfluenza virus. ■ ACUTE non-specific diagnosis of RTI such as sinusitis, pharyngitis, laryngitis. ■ ACUTE non-specific symptoms of RTI such as cough, stridor, shortness of breath, throat pain. ■ EXCLUDES chronic conditions such as chronic bronchitis, asthma without acute exacerbation, chronic sinusitis, allergic conditions.	032.9: Diphtheria, unspecified	2	0.03
		033.0: Whooping cough	39	0.55
		034.0: Streptococcal sore throat and scarlet fever	735	10.38
		460.9: Acute nasopharyngitis (common cold)	20,179	285.05
		465.9: Acute upper respiratory infection, unspecified	223,128	3,151.89
		466.0: Acute bronchitis	62,662	885.16
		480.9: Viral pneumonia, unspecified	837	11.82
		486.9: Pneumonia, organism unspecified	19,755	279.06
		784.1: Throat pain	3,061	43.24
		786.2: Cough	43,291	611.53

[1] Among claims for all 7,079,171 visits to 1,098 participating primary care physician in the province of Quebec in 2005-2007 (32).
[2] Prevalence per 100,000 visits. The prevalence of the diagnostic group 'respiratory syndrome' was 128 per 1,000 claims.

registries,[22,23] patient self-report,[24,25] clinical information systems,[26] medical record review,[27–29] and clinical measurement.[25,30,31] The ecological approach to validating diagnostic groups is inferior because it does not permit the estimation of the sensitivity, specificity, positive predictive value (PPV), or negative predictive value (NPV) of the diagnostic group. In contrast, validation studies based on the direct comparison of individual records identified from 2 different data sources are more informative; however, their design and analysis can pose important challenges. These challenges commonly arise because: (1) the diagnostic group is made up of several diagnostic codes that can have vastly different prevalences in the database; and/or (2) the diagnostic group itself has a low prevalence in the database.

DESCRIPTION OF THE CHALLENGES IN VALIDATING DIAGNOSTIC GROUPS

Challenge 1: Estimating the PPV and NPV of a Diagnostic Group Composed of Several Diagnostic Codes, Each With a Different Population Prevalence

One problem commonly encountered in validation studies of diagnostic groups composed of diagnostic codes in health care utilization data is that each of the diagnostic codes has a different prevalence in the population (example in Table 1). Because of this variation in prevalence, a simple random sample of health care utilization records with diagnostic codes belonging to the diagnostic group may fail to capture enough records with low-prevalence diagnostic codes to generate a reliable estimate of their accuracy. Knowledge of the accuracy of individual diagnostic codes within a diagnostic group is crucial to understanding how a given diagnostic group behaves in different situations and populations (eg, variation in diagnostic coding practices between physician specialty groups and care settings), and to "refine" or modify diagnostic groups to suit different purposes (eg, when a more sensitive vs. specific diagnostic group is needed for disease screening vs. confirmation).

Furthermore, previous validation studies have found that low-prevalence diagnostic codes are more likely to be false positives than higher-prevalence ones.[33,34] Therefore, estimation of the PPV for a diagnostic group must accurately reflect the population prevalence of the individual diagnostic codes in that diagnostic group; oversampling low-prevalence diagnostic codes and failing to adjust for such a sampling strategy in the analysis could lead to an underestimation of the PPV.

Challenge 2: Estimation of the Sensitivity and Specificity of a Diagnostic Group With a Low Population Prevalence

Another challenge in validating diagnostic groups based on diagnostic codes in health care utilization data arises when the prevalence of the diagnostic group itself is low. In such a situation, it may be prohibitively costly, and perhaps infeasible, to use a simple random sample of health care utilization data. With this approach, one would need to review a very large number of records to capture a sufficient number with a diagnostic code that belongs to the diagnostic group. To address this problem, investigators typically sample a larger proportion of claims with a diagnostic code belonging to the diagnostic group of interest than without. For example, a validation of the diagnostic group for "respiratory

syndrome" used for surveillance by both the Centers for Disease Control and Prevention (CDC) and the US Department of Defense,[27] sampled 454 records positive for respiratory syndrome and 2020 records negative for respiratory syndrome, for a sample prevalence of respiratory syndrome of 22.5%; whereas the authors did not report the population prevalence for respiratory syndrome, we estimated it to be nearly half that of the sample prevalence, that is, 12.8%.[33] Failure to account for such a stratified sampling strategy in the statistical analysis can lead to an overestimation of sensitivity and underestimation of specificity,[35] a phenomenon known as verification bias in the context of diagnostic test evaluation.

A correction for verification bias was first described in the context of diagnostic test evaluation by Begg and Greenes.[35] In that context, it occurs when patients first undergo test A (eg, fecal occult blood test), and then, based on the clinician's interpretation of the results from test A in the context of other relevant clinical factors (eg, other signs, symptoms, family history), a subgroup of patients who underwent test A are selected to undergo test B (eg, colonoscopy). Verification bias arises in the estimation of the sensitivity and specificity of test A (the screening test) when a nonrepresentative sample of patients who underwent test A are selected to undergo test B (the "verification" test). Typically, a larger proportion of patients who tested *positive* on test A (the screening test) will be selected to undergo test B (the verification test), as compared with the fraction of patient who tested *negative* on test A who are selected to undergo test B. When estimating the sensitivity and specificity of test A, failure to account for the mechanism whereby patients are selected to undergo test B (the "verification" test) results in verification bias. Consequently, the sensitivity of test A is overestimated and its specificity is underestimated.

Although verification bias was first described in diagnostic test evaluation, it can arise in any validation study that uses a stratified random sampling strategy whereby the proportion of "test A positives" sampled is different from (typically larger than) the proportion of "test A negatives" sampled. In other words, verification bias can arise in any validation study where the prevalence of "test A positives" in the study sample is different from the prevalence of "test A positives" in the study population. Therefore, when we apply this principle to validation studies of diagnostic groups measured in health care utilization databases, we conclude that verification bias can arise whenever the sampled proportion of records *with* a diagnostic code belonging to the diagnostic group is different from the sampled proportion of records *without* such a diagnosis.

In the next sections of this paper, we propose a sequential 2-step approach for validating diagnostic groups based on health care utilization data: step 1, estimating the PPV and NPV of a diagnostic group composed of several diagnostic codes, each code with a different prevalence, and step 2, estimating the sensitivity and specificity of a diagnostic group with low population prevalence. Then, we illustrate our proposed method through application to data from a validation study[33] of the CDC's diagnostic group for surveillance of acute respiratory illness.[32]

METHODS

Notation and study design: The notation we will be using is summarized in Table 2. For each diagnostic code, the entries in the 2×2 table of claims versus medical charts data are denoted using capital letters for the population level (A_m, B_m, C_m, D_m) and small letters for the stratified sample level (a_m, b_m, c_m, d_m). For each diagnostic group, at the population level, the number of positive claims is denoted by A_y+B_y and the number of negative claims by C_y+D_y. The A_y+B_y claims positive for the diagnostic group were first stratified by diagnostic code and then a random sample was drawn from each code for verification with the medical chart. A single random sample was also drawn from the C_y+D_y negative claims for verification, this sampled was frequency-matched to the positive claims on month of visit to avoid seasonal bias. Note that each 2×2 table in Table 2 is created by combining the results of claims which are positive for an individual diagnostic codes and negative for an entire diagnostic group. The entries c_m, d_m are not mentioned explicitly as they are included within c_y, d_y.

Step 1: Estimating the PPV and NPV of a Diagnostic Group Composed of Several Diagnostic Codes, Each With a Different Population Prevalence

When a diagnostic group is composed of diagnostic codes that each have a different population prevalence, we propose stratifying the sample of claims with a diagnosis belonging to the diagnostic group by the individual diagnostic codes that make up the diagnostic group (ie, sampling at the level of the diagnostic code rather than at the level of the diagnostic group).

Using such a stratified sampling strategy, the PPV of individual diagnostic codes can be estimated directly from the 2×2 tables based on the stratified sample using the following usual formula:

$$PPV_m = \frac{a_m}{(a_m+b_m)}, \quad (1)$$

where m denotes an individual diagnostic code within a diagnostic group.

As shown by Begg and Greenes, the PPV of individual diagnostic codes can be estimated without verification bias despite the stratified sampling strategy.[35] It can be shown using Bayes theorem that the PPV of the diagnostic group is the weighted average of the PPV of each diagnostic code within the group, the weight being the estimated prevalence of each diagnostic code in the population:

$$PPV_y = \sum_{m=1}^{n} \left(PPV_m \times \frac{A_m+B_m}{A_y+B_y} \right), \quad (2)$$

where n denotes the total number of diagnostic codes in diagnostic group y. The weight was obtained as the number of visits with a given diagnostic code among the sampled claims, divided by the total number of visits positive for that diagnostic group among all the claims from participating physicians. Note that A_y and B_y are not observed separately due to the fact that not all claims are verified, but their sum, A_y+B_y, is observed.

TABLE 2. Notation Used in the Methods Section and Example of the Data Used in Our Application (Excerpted From Appendix A, Supplemental Digital Content 1, http://links.lww.com/MLR/A876)

Study Population (Total Claims)	Stratified Random Sample (Verified Claims)

NOTATION:

Where the subscript m stands for an ICD-9 code in a diagnostic group

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	ICD-9 code m present	A_m	B_m	$A_m + B_m$
	ICD-9 code not in the diagnostic group	C_m	D_m	$C_m + D_m$
	Total	$A_m + C_m$	$B_m + D_m$	$A_m + B_m + C_m + D_m$

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	ICD-9 code m present	a_m	b_m	$a_m + b_m$
	ICD-9 code not in the diagnostic group	c_y	d_y	$c_y + d_y$
	Total	$a_m + c_y$	$b_m + d_y$	$a_m + b_m + c_y + d_y$

EXAMPLE OF ICD-9 CODES WITHIN THE CDC'S DIAGNOSTIC GROUP FOR 'RESPIRATORY SYNDROME':

465.9 Acute upper respiratory infection of unspecified site

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	465.9 present	?	?	223,128
	ICD-9 code not in the diagnostic group	?	?	6,856,043
	Total	?	?	7,079,171

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	465.9 present	76	8	84
	ICD-9 code not in the diagnostic group	669	3,451	4,120
	Total	745	3,459	4,204

466.0 Acute bronchitis

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	466.0 present	?	?	62,662
	ICD-9 code not in the diagnostic group	?	?	7,016,509
	Total	?	?	7,079,171

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	466.0 present	60	7	67
	ICD-9 code not in the diagnostic group	669	3,451	4,120
	Total	729	3,458	4,187

481.9 Viral pneumonia, unspecified

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	481.9 present	?	?	837
	ICD-9 code not in the diagnostic group	?	?	7,078,334
	Total	?	?	7,079,171

		Chart review (gold standard): Syndrome present	Chart review (gold standard): Syndrome absent	Total
Physician claims	481.9 present	1	0	1
	ICD-9 code not in the diagnostic group	669	3,451	4,120
	Total	670	3,451	4,121

? indicates unknown values.

Using standard statistical theory for stratified samples,[36] the variance of the estimated PPV for the diagnostic group, can be expressed as follows:

$$\sigma^2_{PPV_y} = \sum_{m=1}^{n}\left(\frac{A_m+B_m-(a_m+b_m)}{A_m+B_m}\times\left(\frac{PPV_m\times(1-PPV_m)}{(a_m+b_m)}\right)\times\left(\frac{A_m+B_m}{A_y+B_y}\right)^2\right). \quad (2a)$$

As explained earlier the C_y+D_y "group-negative" records should be negative for *all* "group-positive" diagnostic codes, not only *a single* diagnostic code in the group; for example, "group-negative" records for the CDC's diagnostic group for respiratory syndrome should not include "influenza" or any other acute respiratory infection. Therefore, unlike the PPV, the NPV and its variance are estimated at the level of the diagnostic group, not at the level of individual diagnostic codes, from the 2×2 table in Table 2 using the usual formula:

$$NPV_y = \frac{d_y}{(d_y+c_y)}. \quad (3)$$

$$\sigma^2_{NPV_y} = \frac{(D_y+C_y-(d_y+c_y))}{(D_y+C_y)}\times\frac{NPV_y\times(1-NPV_y)}{(d_y+c_y)}. \quad (3a)$$

Step 2: Estimation of the Sensitivity and Specificity of a Diagnostic Group With a Low Population Prevalence

In most validation studies of low-prevalence diagnostic groups, a larger proportion of records *with* a diagnostic code belonging to the diagnostic group (ie, "group-positive" records) are sampled than records *without* any such diagnosis (ie, "group-negative" records); this is often done to maximize efficiency and minimize cost by validating fewer claims in total.[37] However, failing to take this stratified random sampling strategy into account in the analysis can lead to verification bias: sensitivity is overestimated, specificity is underestimated, and the bias is typically larger for sensitivity than specificity.[35]

A method for correcting for verification bias was published by Begg and Greenes in 1983.[35] It involves taking into account the relative difference in the sampled proportions between the group-positive records and the group-negative records in the estimation of sensitivity and specificity. When the validated claims were randomly sampled within group-positive and group-negative strata, estimation of sensitivity and specificity can be achieved by re-weighting for the different sampling fractions.[35] We propose the following equations to estimate the sensitivity (*Sn*) and specificity (*Sp*) of a diagnostic group *y* using the PPV and NPV estimates derived in the previous section, and the proportion (*p*) of records with a diagnostic code belonging to the diagnostic group,[38] while correcting for verification bias[35]:

$$p_y = \frac{A_y+B_y}{A_y+B_y+C_y+D_y}. \quad (4)$$

$$Sn_y = \frac{PPV_y\times p_y}{\left((PPV_y\times p_y)+(1-NPV_y)\right)\times(1-p_y)}. \quad (5)$$

$$Sp_y = \frac{NPV_y\times(1-p_y)}{\left((NPV_y\times(1-p_y))+(1-PPV_y)\right)\times p_y}. \quad (6)$$

To estimate the variance for sensitivity and specificity, we used the equations that appear in the original paper on verification bias by Begg and Greenes.[35]

APPLICATION

We illustrate the proposed methods by application to a validation study[33] of the CDC's diagnostic group for respiratory syndrome surveillance.[32] In this study, we validated diagnostic codes recorded in reimbursement claims for primary care physician visits; specifically, we compared International Classification of Diseases 9th Revision (ICD-9) codes belonging to the CDC's diagnostic group for respiratory syndrome against diagnoses obtained from chart review for the same patient visit.[33] In brief, we initially selected a random sample of 3600 community-based primary care physicians practicing in the fee-for-service system in the province of Quebec, Canada. We then randomly selected 10 visits per physician from their claims, stratifying on syndrome type and presence, diagnosis, and month. Double-blinded chart reviews were conducted by telephone with consenting physicians to obtain information on patient diagnoses for each sampled visit. The sensitivity, specificity, and PPV of physician claims were estimated by comparison with chart review.

Our final study sample comprised 1098 (12.6%) participating primary care physicians and 10,529 of the 7,079,171 visits for which they submitted fee-for-service claims to the provincial health insurance program in the 2-year period from October 1, 2005 to September 30, 2007. The CDC's diagnostic group for respiratory syndrome includes 171 individual ICD-9 codes; in our final study sample, the prevalence of these individual ICD-9 codes ranged from zero to 3152 per 100,000 primary care visits and the overall population prevalence of the CDC's diagnostic group for respiratory syndrome was 128.3 per 1000 primary care visits.

Example of Step 1: Estimating the PPV and NPV of a Diagnostic Group Composed of Several Diagnostic Codes, Each With a Different Population Prevalence

In our example (see an excerpt of our data in Table 3 or the full data table in online Appendix A, Supplemental Digital Content 1, http://links.lww.com/MLR/A876), the diagnostic group for respiratory syndrome included diagnostic codes with high population prevalence (eg, 465.9—acute upper respiratory infection of unspecified or multiple sites, prevalence of 31.5 per 1000 primary care visits) and diagnostic codes with low population prevalence (eg, 487.0—influenza with pneumonia, prevalence of 0.04 per

TABLE 3. Example of Data from Our Validation Study[33] of the CDC's Diagnostic Group for "Respiratory Syndrome" (Excerpted From Supplemental Digital Content 1, http://links.lww.com/MLR/A876)

ICD-9 Code*	Corresponding Diagnosis	No. Visits in the Population†	Population Prevalence‡	No. True Positives (a_m)	No. False Positives (b_m)	No. False Negatives (c_y)	No. True Negatives (d_y)	PPV
32.9	Diphtheria, unspecified	2	0.03	1	1	669	3451	0.50
33.0	Whooping cough	39	0.55	9	0	669	3451	1.00
34.0	Streptococcal sore throat and scarlet fever	735	10.38	2	0	669	3451	1.00
75.9	Unspecified infectious mononucleosis	2373	33.52	2	0	669	3451	1.00
115.0	Histoplasmosis	1	0.01	0	1	669	3451	0.00
115.9	Histoplasmosis, unspecified	12	0.17	1	4	669	3451	0.20
382.9	Unspecified otitis media	95,165	1344.30	51	17	669	3451	0.75
460.9	Unspecified acute nasopharyngitis	20,179	285.05	11	0	669	3451	1.00
462.9	Unspecified acute pharyngitis	39,695	560.73	24	0	669	3451	1.00
463.9	Unspecified acute tonsillitis	32,676	461.58	15	1	669	3451	0.94
464.0	Acute laryngitis	12,319	174.02	22	5	669	3451	0.81
465.9	Acute upper respiratory infections of unspecified site	223,128	3151.89	76	8	669	3451	0.90
466.0	Acute bronchitis	62,662	885.16	60	7	669	3451	0.90
466.1	Acute bronchiolitis	5,006	70.71	9	1	669	3451	0.90
481.9	Unspecified viral pneumonia	837	11.82	1	0	669	3451	1.00
482.9	Unspecified bacterial pneumonia	2,915	41.18	1	1	669	3451	0.50
484.5	Pneumonia in anthrax	1	0.01	1	0	669	3451	1.00
485.9	Bronchopneumonia, organism unspecified	2,255	31.85	4	0	669	3451	1.00
486.0	Pneumonia, organism unspecified	1	0.01	1	0	669	3451	1.00
486.9	Pneumonia, organism unspecified	19,755	279.06	15	3	669	3451	0.83
487.0	Influenza with pneumonia	278	3.93	1	0	669	3451	1.00
511.9	Unspecified pleural effusion	366	5.17	2	0	669	3451	1.00
769.9	Respiratory distress syndrome	24	0.34	0	1	669	3451	0.00
784.1	Throat pain	3,061	43.24	6	0	669	3451	1.00
786.0	Dyspnea and respiratory abnormalities	24,392	344.56	21	48	669	3451	0.30
786.2	Cough	43,291	611.53	62	7	669	3451	0.90
786.3	Hemoptysis	930	13.14	3	2	669	3451	0.60
786.5	Chest pain	33,582	474.38	46	21	669	3451	0.69
799.1	Respiratory arrest	70	0.99	1	1	669	3451	0.50

*Based on our previously published validation study results; ICD-9 codes that are included in the CDC diagnostic group for respiratory syndrome but were never used by the 1098 primary care physicians during our 2-year study period do not appear in this table. ICD-9 codes that do not appear in this table were not validated due to insufficient or nonexistent use by Quebec primary care physicians during the study period.

†Among claims for all 7,079,171 visits to 1098 participating primary care physician in the province of Quebec in 2005–2007.[33]

‡Prevalence per 100,000 visits. The prevalence of the diagnostic group "respiratory syndrome" was 128 per 1000 claims.

1000 primary care visits). Given that these 2 prevalences differ by several orders of magnitude (10^3), had we taken a simple random sample of 100 visits that met the diagnostic group for respiratory syndrome, we likely would have failed to capture any visit with a diagnosis code of 487.0—influenza with pneumonia. However, the diagnostic code with the lowest prevalence may be more valuable to the investigators (in this case, 487.0—influenza with pneumonia may be a more specific indicator of influenza infection than 465.9—acute upper respiratory infection of unspecified or multiple sites); therefore, validating the diagnosis with the low prevalence may be highly desirable. Therefore, to ensure that our sample contained a sufficient number of visits with rarely used diagnoses to generate stable estimates of those diagnoses' PPV, we stratified our sample by individual diagnostic code.

In the Methods section, we provided Eq. (2) to obtain an estimate of the PPV of a given diagnostic group when some of the diagnostic codes in the diagnostic group are oversampled relative to others. Solving Eq. (2) using the numbers in Supplemental Digital Content 1, http://links.lww.com/MLR/A876, we obtain the PPV estimate for the CDC's diagnostic group for respiratory syndrome (PPV_y) (where the subscript "y" denotes respiratory syndrome):

$$PPV_y = \sum_{i=1}^{n} \left(PPV_m \times \frac{A_m + B_m}{A_y + B_y} \right)$$

$$PPV_y = \left(PPV_{003.2} \times \left(\frac{A_{003.2}+B_{003.2}}{A_y+B_y}\right)\right) + \ldots + \left(PPV_{799.1} \times \left(\frac{A_{799.1}+B_{799.1}}{A_y+B_y}\right)\right)$$

$$PPV_y = \left(0 \times \left(\frac{3}{907,935}\right)\right) + \ldots + \left(0.50 \times \left(\frac{70}{907,935}\right)\right) = 0.85$$

Had we ignored the stratified sampling at the diagnostic code level, and calculated the PPV at the level of the diagnostic group, using $PPV_y = A_y/(A_y+B_y)$, our PPV estimate would have been underestimated at 0.77. Had we computed a simple (unweighted) average of the PPVs of each diagnostic code in the diagnostic group, using $PPV_y = \sum_{m=1}^{n} PPV_m/n$, our PPV estimate would have been ever further underestimated at 0.63.

As mentioned in the Methods section, the NPV of the diagnostic group is conceptualized only at the level of the diagnostic group (not at the level of the diagnostic code) and therefore can be estimated directly from diagnostic group-level data. When we solve Eq. (3) using the numbers in Supplemental Digital Content 1, http://links.lww.com/MLR/A876, we obtain the following NPV estimate for the CDC's diagnostic group for respiratory syndrome: (where the subscript "y" denotes respiratory syndrome)

$$NPV_y = \frac{d_y}{d_y+c_y} = \frac{3451}{(3451+669)} = 0.84.$$

Example for Step 2: Estimation of the Sensitivity and Specificity of a Diagnostic Group With Low Population Prevalence

In the Methods section, we proposed 2 statistical equations to yield estimates of the overall sensitivity (Eq. (5)) and specificity (Eq. (6)) of a given diagnostic group when, due to low population prevalence, "test A positives" are oversampled relative to "test A negatives." When we estimate sensitivity and specificity of respiratory syndrome using Eq. (4–6) together with the data in Supplemental Digital Content 1, http://links.lww.com/MLR/A876, we obtain the following prevalence, sensitivity, and specificity estimate for the CDC definition of respiratory syndrome (where the subscript "*y*" denotes respiratory syndrome):

$$p_y = \frac{A_y+B_y}{A_y+B_y+C_y+D_y} = \frac{907,935}{7,079,181} = 0.1283.$$

In our study population: the prevalence of respiratory syndrome is 128 per 1000 visits

$$Sn_y = \frac{PPV_y \times p_y}{\left((PPV_y \times p_y)+(1-NPV_y)\right) \times (1-p_y)} = \frac{0.85 \times 0.13}{((0.85 \times 0.13)+(1-0.84)) \times (1-0.13)} = 0.44.$$

$$Sp_y = \frac{NPV_y \times (1-y)}{\left((NPV_y \times (1-p_y))+(1-PPV_y)\right) \times p_y} = \frac{0.84 \times (1-0.13)}{(0.84 \times (1-0.13))+(1-0.85) \times 0.13} = 0.97.$$

Had we not adjusted for verification bias, sensitivity would have been overestimated, and specificity would have been underestimated:

$$Sn_y = \frac{a_y}{a_y+c_y} = \frac{803}{803+669} = 0.55.$$

$$Sp_y = \frac{d_y}{d_y+c_y} = \frac{3,451}{3,451+246} = 0.93.$$

It should be noted that the Begg and Greenes correction for verification bias[35] is highly dependent on p, the proportion of records with a diagnostic code belonging to the diagnostic group; the following 2 simulations illustrate this point:

Simulation 1: Decreased Population Prevalence of Respiratory Syndrome

If the study population had been visits to psychiatrists instead of primary care physicians, the prevalence of the CDC's diagnostic group for respiratory syndrome may have been as low as 10 per 1000 visits instead of 128.2 per 1000 visits. Under such conditions, the sensitivity would have been much lower, and the specificity higher:

$$Sn_y = \frac{PPV_y \times p_y}{\left((PPV_y \times p_y)+(1-NPV_y)\right) \times (1-p_y)} = \frac{0.85 \times 0.01}{((0.85 \times 0.01)+(1-0.84)) \times (1-0.01)} = 0.05.$$

$$Sp_y = \frac{NPV_y \times (1-p_y)}{\left((NPV_y \times (1-p_y))+(1-PPV_y)\right) \times p_y} = \frac{0.84 \times (1-0.01)}{(0.84 \times (1-0.01))+(1-0.85) \times 0.01} = 1.00.$$

Simulation 2: Increased Population Prevalence of Respiratory Syndrome

Conversely, if the study population had been visits to pediatric emergency departments, the prevalence of respiratory syndrome could have easily been as high as 200 per 1000 instead of 128.3 per 1000 visits. Under those conditions, the sensitivity would have been higher, and the specificity would have been slightly lower:

$$Sn_y = \frac{PPV_y \times p_y}{\big((PPV_y \times p_y) + (1 - NPV_y)\big) \times (1 - p_y)} = \frac{0.85 \times 0.20}{((0.85 \times 0.20) + (1 - 0.84)) \times (1 - 0.20)} = 0.64.$$

$$Sp_y = \frac{NPV_y \times (1 - p_y)}{\big((NPV_y \times (1 - p_y)) + (1 - PPV_y)\big) \times p_y} = \frac{0.84 \times (1 - 0.20)}{(0.84 \times (1 - 0.20)) + (1 - 0.85) \times 0.20} = 0.96.$$

DISCUSSION

In this paper, we described 2 common challenges in validating diagnostic groups measured from health care utilization data: (1) the diagnostic group's PPV can be underestimated when ignoring the underlying stratified sampling strategy; and (2) the diagnostic group's sensitivity can be overestimated and its specificity underestimated when stratified sampling strategies to improve data collection cost-efficiency by sampling more "group-positive" records relative to "group-negative" records are used. Next, we proposed a 2-step approach for validating diagnostic groups based on health care utilization data: step 1, estimating the PPV and NPV of a diagnostic group composed of several diagnostic codes, each code with a different prevalence, and step 2, estimating the sensitivity and specificity of a diagnostic group with low population prevalence. We then illustrated our proposed methodological approaches by application to a validation study[33] of the CDC's diagnostic group for respiratory syndrome[32] surveillance, and showed how using a stratified sampling strategy without the corresponding statistical adjustments can lead to the underestimation of the PPV and specificity, and the overestimation of the sensitivity.

As we have shown in this paper, failing to recognize and account for challenges intrinsic to the validation of diagnostic groups from health care utilization data can have a substantial impact on inferences drawn from these data. In the context of quality improvement, the overestimation of the sensitivity of a diagnostic group can lead to the underestimation of the frequency or magnitude of the "problem." For example, if 10 cases of complications from diabetes are detected in a given physician practice using a diagnostic group thought to have a sensitivity of 0.95, one can reasonably expect that 10 is the "true" number of cases of diabetic complications in that practice, and, on that basis, one may choose not to invest in interventions to improve diabetes management. However, if the "true" sensitivity of that diagnostic group is 0.30, then the "true" number of diabetic complications in that practice exceeds 10, and an intervention may be desirable (eg, the "true" number of diabetic complications may now be above the threshold at which implementing a given intervention is considered to be cost-effective). Similarly, population surveillance may be greatly affected by the underestimation of the PPV of a diagnostic group secondary to overlooking large differences in the prevalence of individual diagnostic codes within a diagnostic group; because investigating false-positive alerts is very costly, a surveillance system wrongly thought to have a low PPV may not be implemented at all, or worse, an alert it generates may not be acted upon because of the perception that the alert is very likely to be a false positive. In this way, biases in the estimation of the sensitivity, specificity, and PPV of diagnostic groups based on diagnostic codes in health care utilization data can lead to the inefficient and ineffective allocation of limited resources. However, as we have illustrated, it is possible to obtain estimates of validity measures that are free of verification bias by adapting recognized statistical techniques that have been developed in other areas where selection biases arise due to using a cost-effective, stratified sampling strategy.

REFERENCES

1. National Research Council. *The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine)*. Washington, DC: The National Academies Press; 2007.
2. National Research Council. *Digital Infrastructure for the Learning Health System*. Washington, DC: The National Academies Press; 2011.
3. National Research Council. *Digital Data Improvement Priorities for Continuous Learning in Health and Health Care: Workshop Summary*. Washington, DC: The National Academies Press; 2013.
4. Aguilar-Gaxiola S, Ahmed S, Franco Z, et al. Towards a unified taxonomy of health indicators: academic health centers and communities working together to improve population health. *Acad Med*. 2014;89:564–572.
5. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med*. 2010;363:501–504.
6. Freeman JL, Zhang D, Freeman DH, et al. An approach to identifying incident breast cancer cases using Medicare claims data. *J Clin Epidemiol*. 2000;53:605–614.
7. Tirschwell DL, Longstreth WT. Validating administrative data in stroke research. *Stroke*. 2002;33:2465–2470.
8. Solberg LI, Engebretson KI, Sperl-Hillen JM, et al. Are claims data accurate enough to identify patients for performance measures or quality improvement? The case of diabetes, heart disease, and depression. *Am J Med Qual*. 2006;21:238–245.
9. Birman-Deych E, Waterman AD, Yan Y, et al. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care*. 2005;43:480–485.
10. Best WR, Khuri SF, Phelan M, et al. Identifying patient preoperative risk factors and postoperative adverse events in administrative databases: Results from the Department of Veterans Affairs National Surgical Quality Improvement Program. *J Am Coll Surg*. 2002;194:257–266.
11. Quan HD, Parsons GA, Ghali WA. Validity of information on comorbidity derived from ICD-9-CCM administrative data. *Med Care*. 2002;40:675–685.
12. Humphries KH, Rankin JM, Carere RG, et al. Co-morbidity data in outcomes research—are clinical data derived from administrative databases a reliable alternative to chart review? *J Clin Epidemiol*. 2000;53:343–349.
13. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol*. 1992;45:613–619.

14. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *J Clin Epidemiol*. 2004;57:131–141.
15. Malenka DJ, Mclerran D, Roos N, et al. Using administrative data to describe casemix—a comparison with the medical record. *J Clin Epidemiol*. 1994;47:1027–1032.
16. Starfield B, Weiner J, Mumford L, et al. Ambulatory care groups—a categorization of diagnoses for research and management. *Health Serv Res*. 1991;26:53–74.
17. Mitchell JB, Bubolz T, Paul JE, et al. Using Medicare claims for outcomes research. *Med Care*. 1994;32:JS38–JS51.
18. Virnig BA, McBean M. Administrative data for public health surveillance and planning. *Ann Rev Public Health*. 2001;22:213–230.
19. Poissant L, Taylor L, Huang A, et al. Assessing the accuracy of an inter-institutional automated patient-specific health problem list. *BMC Med Inform*. 2010;10:1–10.
20. Hripcsak G, Soulakis ND, Li L, et al. Syndromic surveillance using ambulatory electronic health records. *J Am Med Inform Assoc*. 2009;16: 354–361.
21. Caudle JM, van Dijk A, Rolland E, et al. Telehealth Ontario detection of gastrointestinal illness outbreaks. *Can J Public Health*. 2009;100: 253–256.
22. Kostylova A, Swaine B, Feldman D. Concordance between childhood injury diagnoses from two sources: an injury surveillance system and a physician billing claims database. *Inj Prev*. 2005;11:186–190.
23. McClish DK, Penberthy L, Whittemore M, et al. Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. *Am J Epidemiol*. 1997;145:227–233.
24. Ostbye T, Taylor DH, Clipp EC, et al. Identification of dementia: agreement among national survey data, medicare claims, and death certificates. *Health Serv Res*. 2008;43:313–326.
25. Muhajarine N, Mustard C, Roos LL, et al. Comparison of survey and physician claims data for detecting hypertension. *J Clin Epidemiol*. 1997;50:711–718.
26. Jollis JG, Ancukiewicz M, Delong ER, et al. Discordance of databases designed for claims payment versus clinical information-systems—implications for outcomes research. *Ann Internal Med*. 1993;119: 844–850.
27. Betancourt JA, Hakre S, Polyak CS, et al. Evaluation of ICD-9 codes for syndromic surveillance in the electronic surveillance system for the early notification of community-based epidemics. *Mil Med*. 2007;172: 346–352.
28. Chapman WW, Dowling JN, Wagner MM. Generating a reliable reference standard set for syndromic case classification. *J Am Med Inform Assoc*. 2005;12:618–629.
29. Losina E, Barrett J, Baron JA, et al. Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. *J Clin Epidemiol*. 2003;56:515–519.
30. Lix LM, Yogendran MS, Leslie WD, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases *J Clin Epidemiol*. 2008;61:1250–1260.
31. van Walraven C, Austin PC, Manuel D, et al. The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model. *J Clin Epidemiol*. 2010;63:1332–1341.
32. Centers for Disease Control and PreventionSyndrome definitions for diseases associated with critical bioterrorism-associated agents. October 23, 2003. Available at: http://www.bt.cdc.gov/surveillance/syndromedef/.
33. Cadieux G, Buckeridge DL, Jacques A, et al. Accuracy of syndrome definitions based on diagnoses in physician claims. *BMC Public Health*. 2011;11:17–26.
34. MacIntyre CR, Ackland MJ, Chandraraj EJ, et al. Accuracy of ICD-9-CM codes in hospital morbidity data, Victoria: implications for public health research. *Aust N Z J Public Health*.; 1997:477–482.
35. Begg CB, Greenes RA. Assessment of diagnostic-tests when disease verification is subject to selection bias. *Biometrics*. 1983;39: 207–215.
36. Cochran WG. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons Inc.; 1977.
37. Irwig L, Glasziou PP, Berry G, et al. Efficient study designs to assess the accuracy of screening-tests. *Am J Epidemiol*. 1994;140:759–769.
38. Kelly H, Bull A, Russo P, et al. Estimating sensitivity and specificity from positive predictive value, negative predictive value and prevalence: application to surveillance systems for hospital-acquired infections. *J Hosp Infect*. 2008;69:164–168.

